Escuela Politécnica Superior

Departamento de Tecnología Electrónica y de las Comunicaciones

CONTRIBUTIONS TO ROBUST PEOPLE DETECTION IN VIDEO-SURVEILLANCE

PhD Thesis written by
Álvaro García Martín
under the supervision of
Prof. José María Martínez Sánchez

Madrid, June 2013

Department: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

PhD Thesis: Contributions to robust people detection in video-surveillance

Author: **Álvaro García Martín**
Ingeniero de Telecomunicación
(Universidad Autónoma de Madrid)

Supervisor: **Jose María Martínez Sánchez**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid , Spain

Year: 2013

Committee: President: **Thomas Sikora**
Technisce Universität Berlin, Germany

Secretary: **Jesús Bescós Cano**
Universidad Autónoma de Madrid, Spain

Vocal 1: **Rita Cucchiara**
Università degli Studi di Modena e Reggio Emilia, Italy

Vocal 2: **Andrea Cavallaro**
Queen Mary University of London, UK

Vocal 3: **Jordi Gonzalez Sabaté**
Universidad Autònoma de Barcelona, Spain

The work described in this Thesis was carried out within the Video Processing and Understanding Lab at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2008 to 2013). It was partially supported by the Universidad Autónoma de Madrid ("FPI-UAM: Programa propio de ayudas para la Formación de Personal Investigador"), by Cátedra Infoglobal-UAM for "Nuevas Tecnologías de video aplicadas a la seguridad" and by the Spanish Goverment (TEC2007-65400 SemanticVideo and TEC2011-25995 EventVideo).

Part of the work related to the motion person model was done while visiting the School of Computer Science at the Carnegie Mellon University (USA) under the supervision of Prof. Alexander G. Hauptmann from September 15 to December 14, 2010. In addition, part of the work related to the people-background segmentation was done while visiting the Multimedia & Vision Research Group at the Queen Mary University of London (UK) under the supervision of Prof. Andrea Cavallaro from September 15 to December 14, 2011.

Todos los caminos del mundo llevan hasta el corazón del guerrero;

él se zambulle sin vacilar en el río de las pasiones que siempre corre por su vida.

-Paulo Coelho (1947-)-

Acknowledgments

This thesis has been an adventure that I have shared with many people. I want to thank all of them who have made the happy ending possible. Each one has contributed to finish this adventure in some way.

First of all, I would like to express my gratitude to my supervisor, Dr. José María Martínez Sánchez. This work could not have been done without his great advice and dedication. I really appreciate his constant support and patience. I would also like to specially thank Dr. Alex Hauptmann from the Carnegie Mellon University and Dr. Andrea Cavallaro from the Queen Mary University for their support, advice and warm hospitality during my short stay under their supervision. I want to thank all three for their contributions to this thesis and for teaching me three different ways of approaching research.

I want to thank especially Dr. Jesús Bescós Cano for his advice and suggestions during the past years and all the rest present and past members of VPU-Lab (Álvaro Bayona, Álvaro García, Fabrizio, Fernando, Javi, Luis Caro, Luis Herranz, Luis Jaime, Luis Salgado, Miguel Angel, Rafa, Santiago, Víctor Fernández and Víctor Valdés). A very special thanks to Marcos who has given me his advice and friendship from the very first day of my college experience.

Last but not least, this work would not have been possible without the continued support provided by my family and all my friends. Amongst them, many thanks to Virginia for making the whole adventure much more bearable and Miguel for giving me the best of himself in those times when I needed it most.

Álvaro García Martín

June 2013

Abstract

Computer vision is a field with multiple lines of research and different application domains, being video surveillance one of the most developed during the last years. During the past years, automatic video surveillance systems have experienced a great development driven by the growing need of security. These automatic systems include several image and video processing techniques for monitoring purposes. Among the different video surveillance tasks, the main objective of this thesis has been the exploration of the state of the art in people detection, analyze the most representative approaches, identify their weaknesses and propose contributions to improve current people detection state of the art.

The people detection task consists mostly of, firstly, the design and training of a person model based on characteristic parameters (motion, dimensions, silhouette, etc) and, secondly, the adjustment of this model to the candidate objects in the scene. Thus, the critical tasks in any people detection algorithm are the generation or extraction of the initial object hypotheses to be people from the scene and the person model used to classify those initial object hypotheses. Firstly, in order to analyze the people detection problems in surveillance scenarios the critical tasks in any people detection algorithm have been identified and a consequently framework for their evaluation have been designed. Secondly, three different people detection algorithms have been proposed and compared with the state of the art, covering all the people detection issues previously identified. Finally, two different people detection post-processing subtasks focused on improving the final detection results have been also proposed.

The performance of the proposed people detection algorithms and post-processing subtasks has been thoroughly evaluated on the proposed evaluation dataset. The experiments conducted demonstrated the advantages and disadvantages of every proposed people detection approach in typical surveillance scenarios. Finally, the inclusion of the proposed post-processing subtasks provides robustness and improves the final detection results.

Resumen

La visión por computador es un campo con múltiples líneas de investigación y diferentes dominios de aplicación, siendo la videovigilancia uno de los más desarrollados en los últimos años. Durante los últimos años, los sistemas de videovigilancia automáticos han experimentado un gran desarrollo empujados por la creciente necesidad de seguridad. Estos sistemas automáticos incluyen múltiples técnicas de procesado de imagen y video con propósitos de monitorización. Dentro de las diferentes tareas que engloba la videovigilancia, el principal objetivo de esta tesis ha sido la exploración del estado del arte de detección de personas, analizar las aproximaciones más representativas, identificar sus debilidades y proponer contribuciones que mejoren el estado del arte de detección de personas.

La detección de personas consiste principalmente, en primer lugar, el diseño y entrenamiento de un modelo de persona basado en parámetros característicos (movimiento, dimensiones, silueta, etc) y, en segundo lugar, el ajuste de este modelo a los objetos candidatos en la escena. Por lo tanto, las tareas críticas de cualquier algoritmo de detección de personas son la generación o extracción de los objetos inicialmente candidatos a ser personas y el modelo de persona usado para clasificar dichos objetos inicialmente candidatos. En primer lugar, con el objetivo de analizar los problemas inherentes a la detección de personas en escenarios de videoseguridad, se han identificado las tareas críticas en cualquier algoritmo de detección de personas y se ha diseñado en consecuencia un marco de trabajo para su evaluación. En segundo lugar, se han propuesto tres algoritmos diferentes de detección de personas y se han comparado con el estado del arte, abarcando todos los problemas previamente identificados que conllevan la detección de personas. Finalmente, también se han propuesto dos algoritmos diferentes de post-procesado orientados a mejorar los resultados finales de detección.

El rendimiento de los algoritmos de detección de personas y de post-procesado propuestos ha sido evaluado exhaustivamente sobre el dataset de evaluación propuesto. Los experimentos realizados demuestran las ventajas e inconvenientes de cada uno de los algoritmos de detección de personas propuesto en escenarios típicos de videovigilancia. Finalmente, la inclusión de los algoritmos de post-procesado propuestos añade robustez y mejora los resultados finales de detección.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Motivation

Computer vision is a field that includes methods for acquiring, processing, analyzing and understanding images; in general, high-dimensional data from the real world in order to produce numerical or symbolic information, e.g., in the forms of decisions. As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner.

Computer vision is an evolving field during the last years with multiple lines of research and different application domains. Video surveillance is one of the most developed domains during the last 10 years [Platanioitis and Regazzoni, 2005; Valera and Velastin, 2005; Haering et al., 2008; Regazzoni et al., 2010]. Video surveillance systems try to extract automatically information from the video sequence and to generate a scene description useful for human interactions with the system: alarms, logs, statistics, indexing and retrieval, etc. The need for providing security to people and their properties in the entire world explains the huge development and expansion of video surveillance systems nowadays.

Within the computer vision field, particularly in the research area of digital image and video processing, there exists a rich variety of algorithms for foreground segmentation, object detection, event recognition, etc, which are being used in surveillance systems. People detection is one of the most challenging problems in this field. The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of views and interactions among different people and objects. This complexity is even higher in real world scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability.

In addition, people detection has a wide range of applications including video surveillance

but also intelligent systems (robotic), image and video indexing, driver assistance systems, video games, etc.

1.2 Objectives

The main objective of this thesis is to explore the state of the art in people detection in surveillance scenarios, analyze the most representative approaches, identify their weaknesses and propose contributions to improve current people detection state of the art.

For achieving this objective, we propose to study two main areas:

- People detection benchmarking:
  - We explore the critical factors applied to the generation of a corpus (dataset and associated ground-truth) and the definition of a performance evaluation methodology, for the evaluation of people detection algorithms in video sequences.
- People detection approaches:
  - People detection based on segmentation vs. exhaustive search. We explore the combination of both techniques, trying to leverage their strengths and address its drawbacks. Firstly, the initial objects candidates to be person can be extracted using segmentation and, then, those selected candidates can be processed with an exhaustive search. The preliminary segmentation reduces the exhaustive search critical factor, eliminating easily most of the false negative examples. And the subsequent exhaustive search over the already detected objects reduces the segmentation dependency, but being still robust to partial detections and overlapping objects.
  - People detection based on appearance vs. motion. We explore the combination of appearance and motion information. Most of the existing approaches are based only on appearance information due to the fact that appearance provides a much more discriminant information about people detection. Human appearance varies due to environmental factors such as light conditions, clothing, contrast, etc, apart from the huge intrinsic people variability such as different heights, widths, poses, etc. However, the motion information is independent of all these factors.
  - Detection-by-Tracking and Tracking-by-Detection. The benefits of using the detection information to improve the tracking have been already reported in the state of the art, but we also introduce the opposite flow of information to improve the detection, so that we define a collaborative scheme system that integrates the people detection and tracking information into a single system and improves both tasks simultaneously.

– People detection post-processing based on novel people-background segmentation. While the focus of person detection approaches is to obtain a high detection performance and to reduce false positive detections, we aim at determining the areas without people in the scene by giving a higher penalty to pixels representing a person, but that have been incorrectly classified as background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with bias on people. People-background segmentation gives us information about where there are not people in the scene. We can use this information to eliminate, or at least reduce, the number of false positives and, therefore, improve the global detection results.

– Decision-level fusion of people detectors. We explore the combination at decision-level of multiple people detectors from the state of the art in order to take advantage of their independent strengths and at the same time reduce their drawbacks and limitations, and, therefore, improve the global detection performance in typical video surveillance environments.

1.3 Major contributions

The significant novel contributions of this thesis are summarized below:

1. A complete framework for the evaluation of people detection algorithms under different complexity conditions. It includes a more complete people detection corpus in surveillance scenarios than the ones available in the state of the art and a performance evaluation methodology.

2. A robust people detector based on appearance information that is capable of operating in real time in low and medium complexity scenarios.

3. A people detection approach based on motion and the combination of appearance and motion information in order to solve people detection in more complex and realistic scenarios.

4. A detection/tracking collaborative scheme that integrates detection (based on appearance and motion) and tracking information. The collaborative system consists of successive stages of mutual information exchange, so that the improvement introduced by one process becomes a potential self-improvement in the following stages, improving the detection results over time in complex and realistic scenarios.

5. A people-background segmentation approach that aims to ensure that there are no people or body parts assigned to the background class at the cost of potentially increasing the number of background pixels classified as people.

6. A people detection post-processing subtask. It make use of the information about where there are not people in the scene obtained with the people-background segmentation; reducing the number of false positives, but maintaining, as much as possible, the number of positive detections.

7. The combination or fusion of independent people detectors from the state of the art at decision-level. It includes a multi people detection combination criteria and the application of traditional fusion techniques: average, product, minimum, maximum and median.

1.4 Structure of the document

This document is structured in three parts and appendixes, which are organized as follows:

- Part I: Introduction
  - *Chapter 1: Introduction.* This chapter presents the motivation, the objectives, the main contributions and the structure of this thesis.
  - *Chapter 2: State of the art.* This chapter describes an overview of the state of the art in automatic people detection in video sequences.
  - *Chapter 3: People detection benchmarking framework.* This chapter proposes a corpus (dataset and associated ground-truth) and defines the performance evaluation methodology, for the evaluation of people detection algorithms in video sequences.
- Part II: People detection approaches
  - *Chapter 4: Real time people detection based on appearance information.* This chapter proposes a real time people detection approach.
  - *Chapter 5: People detection based on appearance and motion information.* This chapter proposes a new people detection approach based on motion and the combination of appearance and motion information.
  - *Chapter 6: Collaborative people detection and tracking.* This chapter proposes a collaborative people detection and tracking system.
  - *Chapter 7: People detection using people-background segmentation confidence.* This chapter proposes a people-background segmentation approach and a new people detection post-processing subtask based on this people-background segmentation.
  - *Chapter 8:Decision-level fusion of people detectors.* This chapter proposes the combination at decision-level of multiple people detectors.
- Part III: Conclusions

- *Chapter 9: Achievements, conclusions and future work.* It concludes this document summarizing the main results and future research lines.

- IV: Appendixes

  - *Appendix A: People detectors.* This appendix presents a brief introduction of the different people detection approaches used from the state of the art.
  - *Appendix B: People-background segmentation experimental results.* This appendix describes the experimental evaluation of the proposed people-background segmentation approach.
  - *Appendix C: Decision-level fusion of people detectors additional experimental results.* This appendix describes additional experimental results of the proposed combination at decision-level of multiple people detectors.
  - *Appendix D: Publications.*
  - *Appendix E: Logros, conclusiones y trabajo futuro.*

A representative diagram of the contents of the thesis is depicted in Figure 1.1.

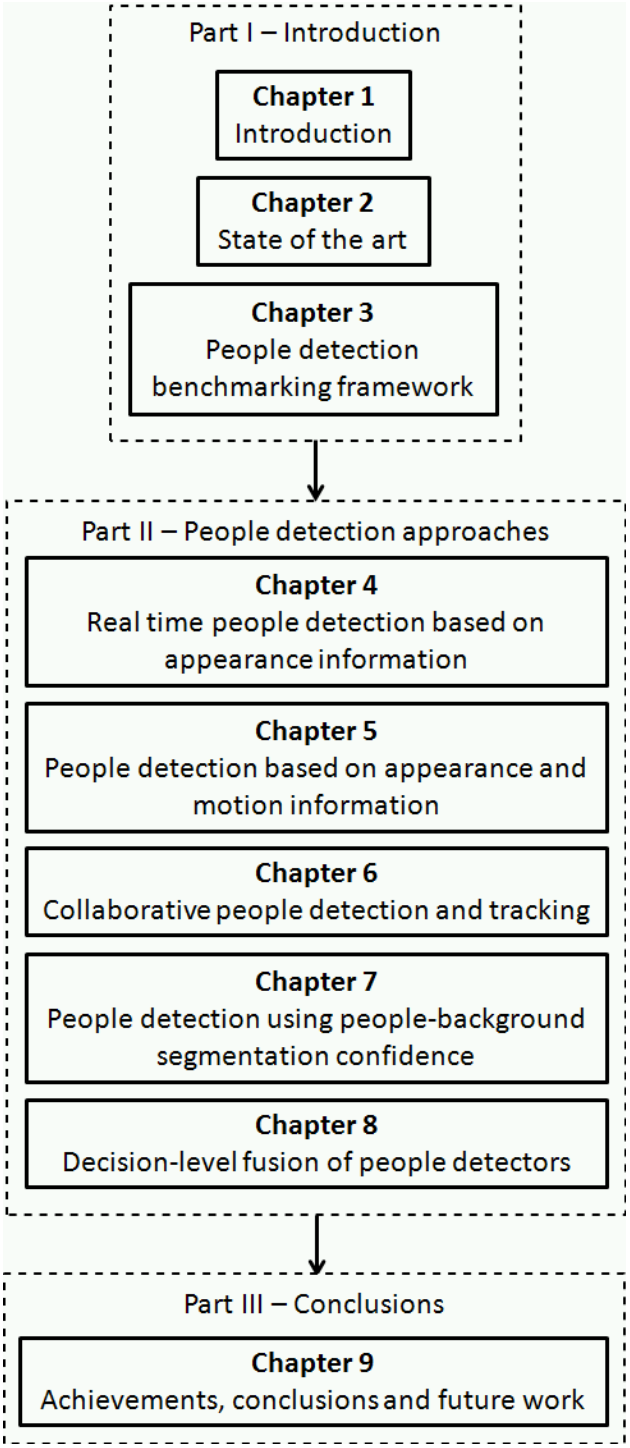


Fig. 1.1. Diagram of the contents of the thesis.

Chapter 2

State of the art

2.1 Introduction

Automatic people detection in video sequences [Enzweiler and Gavrila, 2009; Gerónimo et al., 2010; Dollár et al., 2012] is one of the most challenging problems in computer vision. The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of view and interactions among different people and objects. This complexity is even higher in typical real world surveillance scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability.

In this chapter, we will make an overview of the state of the art in automatic people detection in video sequences. Firstly, section 2.2 presents a brief review of previous surveys from the state of the art and section 2.3 describes the basic architecture of every people detector surveillance system. Then, the proposed classification of state of the art people detection algorithms is described in section 2.4. Finally, section 2.5 summarizes the chapter with some conclusions.

2.2 Survey of the state of the art

There is a large number of people detection surveys in the literature, some of them cover only partially the state of the art or are clearly focused on some particular video surveillance application. [Enzweiler and Gavrila, 2009] presents a survey of people detection and also the integration of the detectors into full systems. It decomposes people detection approaches into three processing tasks: generation of initial object hypotheses or Regions of Interest (ROI) selection, verification (classification) and temporal integration (tracking). [Gerónimo et al., 2010] also presents a survey of people detection, but with a clear focus on driver assistance systems and defines a processing pipe line: preprocessing, foreground segmentation, object classification, verification or refinement, tracking and application. [Dollár et al., 2012] presents an overview of people

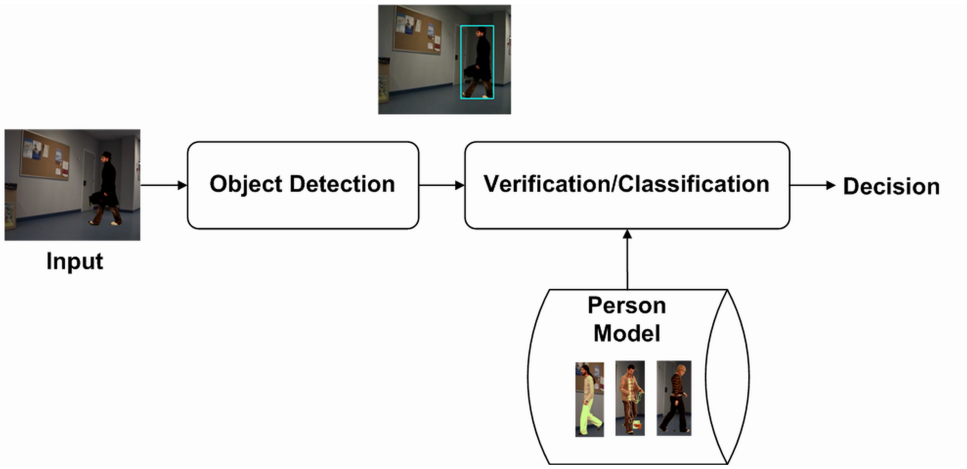


Fig. 2.1. Canonical people detection architecture.

detection algorithms focused only on sliding window approaches.

In our case we are focused on identifying and classifying all the different approaches from the state of the art, regardless of their subsequent video surveillance application. We decompose the people detection in subtasks, identify the critical tasks and classify the state of the art according to these critical tasks. In this way, we are able to analyze the strengths and weaknesses of each approach independently and for each critical task. Any other possible additional subtask is considered as a specific video surveillance application preprocessing or post-processing subtask and they are not part of the main scope of this review.

2.3 Architecture of people detection systems

As defined for surveillance canonical systems [Valera and Velastin, 2005; Hu et al., 2004], every people detection approach consists mostly of, firstly, the design and training (if training is required) of a person model based on characteristic parameters (motion, dimensions, silhouette, etc) and, secondly, the adjustment of this person model to the candidates to be person in the scene. All candidates that adjust to the model will be detected or classified as person, whilst all the others will not be detected neither classified as person. Figure 2.1 shows the basic architecture of any people detector.

Input

There are many different possible input formats, which determine the type of input information available to the detector. In relation to computer vision, the basic processing input unit is the image or the frame in the case of video processing. Input images can be of multiple resolutions, 2D or 3D, color or gray scale, visible or infrared spectrum, etc. Input videos can be from static or mobile cameras, mono or stereo-vision, etc.

Object detection

Object detection consists in the generation or extraction from the scene of the initial object hypotheses, that is, candidates to be a person. This is a critical task for people detection. The chosen approach (e.g., background subtraction, sliding-window) will be very determinant for some global detection performance factors: processing speed, detection results, robustness to scene variations, etc.

Person model

The person model defines the characteristics and rules that the objects must meet in the scene in order to be considered as people. Like the previous step, this is also a critical task for people detection. The chosen approach (e.g., holistic, part-based) will be very determinant in some global detection performance factors: processing speed, robustness to pose variations, partial occlusions, etc.

Verification or Classification

The verification or classification task can be considered as a standard pattern recognition issue. This process compares previously trained object models and the generated object model from an image or sequence.

Decision

According to the comparison or similarity calculated in the previous stage, a final decision must be taken. Depending on the subsequent application, the decision may be binary (person or no person) or fuzzy (a confidence value or probability of being a person).

2.4 Proposed classification of state of the art people detection

This section describes the proposed classification of people detection algorithms carried out and describes the different representative people detection algorithms selected from the state of the art[1]. Many criteria can be used to classify people detection algorithms; for example, the techniques used (e.g., background or foreground extraction, movement estimation or compensation), the type of models used (e.g., stick figure-based, statistical, movement), the use of 2D or 3D information, the sensor modality (e.g., visible light, infra-red), the sensor multiplicity (monocular, stereo or multicamera), the sensor placement (centralized vs. distributed), the sensor mobility (stationary vs. moving), etc.

As already mentioned in the previous section, the two main critical tasks of people detection (object detection and person model) determine the global detection performance; therefore, it has been decided to propose a classification of the state of the art algorithms according to these tasks. In the remainder of this section, we describe the classification of different algorithms

[1] Any classification system could be perfectly debated because it depends on the discriminative aspects on which its hierarchy is based.

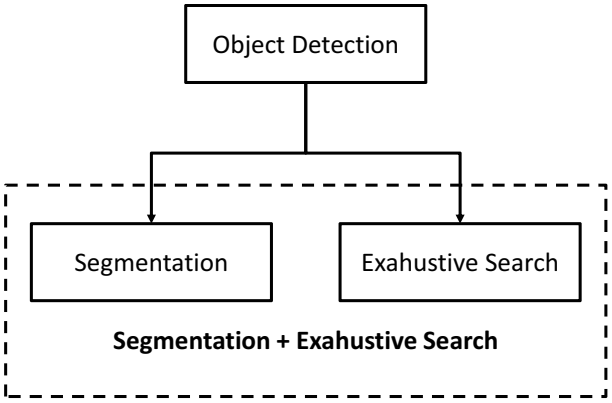


Fig. 2.2. People detection classification I.

from the state of the art. Firstly, we classify the people detection algorithms according to the approach used to generate or extract the initial candidate objects to be a person, whilst the second classification is based on the chosen person model (see Table 2.1).

2.4.1 Object detection approach or Initial object hypotheses

There are two main conventional object detection approaches (see Figure 2.2): those based on some kind of segmentation of the scene in foreground (objects) and background [Cutler and Davis, 2000; Haritaoglu et al., 2000; Sprague and Luo, 2002; Xu and Fujimura, 2003; Giebel et al., 2004; Zhao and Nevatia, 2004; Zhou and Hoang, 2005; Harasse et al., 2006; Hussein et al., 2006; Gavrila and Munder, 2007; Koenig, 2007; Fernández-Carbajales et al., 2008; Kilambi et al., 2008] and those based on an exhaustive scanning approach [Viola et al., 2003; Leibe and Schiele, 2004; Okuma et al., 2004; Sidenbladh, 2004; Viola and Jones, 2004; Dalal and Triggs, 2005; Wu and Nevatia, 2005; Dalal and Triggs, 2006; Seemann and Schiele, 2006; Zhu et al., 2006; Avidan, 2007; Cui et al., 2007; Leibe et al., 2007; Wu and Nevatia, 2007; Zhang et al., 2007; Andriluka et al., 2008; Leibe et al., 2008; Li et al., 2008; Ren, 2008; Wojek et al., 2008; Andriluka et al., 2009; Ess et al., 2009; Breitenstein et al., 2010; Felzenszwalb et al., 2010; Stalder et al., 2010; Yu et al., 2011]. There are also some approaches that try to combine both approaches together[Alonso et al., 2007]. In any case, the result of this stage is the location and dimension (bounding box or blob[2]) of the different objects in the scene candidates to be a person. Table 2.2 summarizes the different approaches from the state of the art according to the used object detection approach.

[2]In the literature, both terms have been used without any distinction, for the rest of the thesis we also use both without any distinction.

Object Detection		Person Model		
		Motion	Appearance	
			Holistic	Part-based
	Segmentation	[Cutler and Davis, 2000; Giebel et al., 2004]	[Xu and Fujimura, 2003; Giebel et al., 2004; Zhao and Nevatia, 2004; Zhou and Hoang, 2005; Hussein et al., 2006; Gavrila and Munder, 2007; Koenig, 2007; Fernández-Carbajales et al., 2008; Kilambi et al., 2008]	[Haritaoglu et al., 2000; Sprague and Luo, 2002; Harasse et al., 2006; Alonso et al., 2007]
	Exhaustive Search	[Viola et al., 2003; Okuma et al., 2004; Sidenbladh, 2004; Dalal and Triggs, 2006; Avidan, 2007; Cui et al., 2007; Leibe et al., 2007; Wu and Nevatia, 2007; Andriluka et al., 2008; Li et al., 2008; Ren, 2008; Ess et al., 2009; Breitenstein et al., 2010; Stalder et al., 2010; Yu et al., 2011]	[Viola et al., 2003; Leibe and Schiele, 2004; Okuma et al., 2004; Viola and Jones, 2004; Dalal and Triggs, 2005, 2006; Seemann and Schiele, 2006; Zhu et al., 2006; Avidan, 2007; Cui et al., 2007; Leibe et al., 2007; Zhang et al., 2007; Leibe et al., 2008; Li et al., 2008; Ren, 2008; Wojek et al., 2008; Ess et al., 2009; Breitenstein et al., 2010; Stalder et al., 2010; Yu et al., 2011]	[Wu and Nevatia, 2005; Alonso et al., 2007; Wu and Nevatia, 2007; Andriluka et al., 2008, 2009; Ess et al., 2009; Felzenszwalb et al., 2010]

Table 2.1: State of the art people detection classification.

Approach	Segmentation	Exhaustive search
[Cutler and Davis, 2000; Haritaoglu et al., 2000; Zhao and Nevatia, 2004; Zhou and Hoang, 2005; Hussein et al., 2006; Fernández-Carbajales et al., 2008; Kilambi et al., 2008]	Background subtraction	-
[Sprague and Luo, 2002; Harasse et al., 2006]	Color information	-
[Xu and Fujimura, 2003; Giebel et al., 2004; Gavrila and Munder, 2007; Koenig, 2007]	3D information	-
[Alonso et al., 2007]	3D information	Bounded sliding-window
[Viola et al., 2003; Okuma et al., 2004; Sidenbladh, 2004; Viola and Jones, 2004; Dalal and Triggs, 2005; Wu and Nevatia, 2005; Dalal and Triggs, 2006; Zhu et al., 2006; Avidan, 2007; Cui et al., 2007; Wu and Nevatia, 2007; Zhang et al., 2007; Li et al., 2008; Ren, 2008; Wojek et al., 2008; Breitenstein et al., 2010; Felzenszwalb et al., 2010; Stalder et al., 2010; Yu et al., 2011]	-	Sliding-window
[Ess et al., 2009]	-	Sliding-window or Feature-based
[Leibe and Schiele, 2004; Seemann and Schiele, 2006; Leibe et al., 2007; Andriluka et al., 2008; Leibe et al., 2008; Andriluka et al., 2009]	-	Feature-based

Table 2.2: State of the art people detection classification I.

2.4.1.1 Segmentation

Image segmentation is often used to partition an image into separate regions, which ideally correspond to different real world objects. More precisely, it is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristic or computed property, such as color, motion, intensity, texture, etc. Adjacent regions must be significantly different with respect to the same characteristic. The ideal final result is to locate and discriminate objects in the scene (foreground) vs. the rest of the image (background).

Currently, there are many approaches from the state of the art that use some kind of segmentation as a preliminary step in the people detection task. In particular, the use of background subtraction is very popular in surveillance applications [Cutler and Davis, 2000; Haritaoglu et al., 2000; Zhao and Nevatia, 2004; Zhou and Hoang, 2005; Hussein et al., 2006; Fernández-Carbajales et al., 2008; Kilambi et al., 2008]. They try to detect moving objects from the difference between the current frame and a reference frame (background model) and threshold the results to generate the objects of interest. There are some approaches that use color segmentation [Sprague and Luo, 2002; Harasse et al., 2006], owing the fact that the skin color facilitates the people segmentation and detection process. There are multiple approaches that use some kind of 3D information to facilitate the segmentation by stereo-vision [Giebel et al., 2004; Alonso et al., 2007; Gavrila and Munder, 2007] or directly with 3D cameras [Xu and Fujimura, 2003; Koenig, 2007].

In relation to people detection, the use of segmentation directly generates the objects candidates to be a person and rejects easily irrelevant areas of the image, i.e., without objects of interest. For this reason, the subsequent classification task is clearly simplified and, therefore, the person model usually is simpler and has lower computational cost. However, as there is a strong dependence with the segmentation, all the segmentation problems are inherited (under and over segmentation). These problems can affect the global detection performance, mainly limiting the maximum detection rate (undetected objects), but also increasing the number of false detections (partial object detections or overlapping objects). Furthermore, these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation.

2.4.1.2 Exhaustive search

The other technique to obtain initial object location hypotheses is the exhaustive search. Usually, it consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations. Through this mechanism a dense detection confidence map or volume (scale and location) is obtained; in order to arrive at individual detections, these approaches must search for local maxima in the density volume and, then, apply some form of non-maximum suppression. There are many people detection approaches from the state of the art that use this technique, in fact, this technique is currently the most widely used. Within

this technique, two different approaches can be used as stated in [Breitenstein et al., 2010]. On one hand, there are some approaches that obtain this density volume implicitly sampling in a discrete 3D grid (location and scale) by evaluating different detection windows with a classifier; this is the case of using sliding-window based detectors such as [Viola et al., 2003; Okuma et al., 2004; Sidenbladh, 2004; Viola and Jones, 2004; Dalal and Triggs, 2005; Wu and Nevatia, 2005; Dalal and Triggs, 2006; Zhu et al., 2006; Alonso et al., 2007; Avidan, 2007; Cui et al., 2007; Wu and Nevatia, 2007; Zhang et al., 2007; Li et al., 2008; Ren, 2008; Wojek et al., 2008; Ess et al., 2009; Breitenstein et al., 2010; Felzenszwalb et al., 2010; Stalder et al., 2010; Yu et al., 2011]. On the other hand, there are some approaches that create this density volume explicitly in a bottom-up fashion through probabilistic votes cast by local features matching; this is the case of using feature-based detectors such as [Leibe and Schiele, 2004; Seemann and Schiele, 2006; Leibe et al., 2007; Andriluka et al., 2008; Leibe et al., 2008; Andriluka et al., 2009; Ess et al., 2009].

Generally, those detectors that use this kind of approaches are more robust to scale and pose variations and, therefore, more reliable in complex environments than those based on segmentation. However unlike the previous case, the classification task is not simplified, it is even more complex because the person model must be able to classify correctly a great number of negative examples (potential false positive detections). In addition to the increased person model complexity, the exhaustive search process itself usually requires a higher computational cost, which makes difficult to fulfill real time requirements. Although some proposals have studied this problem [Zhu et al., 2006; Zhang et al., 2007; Wojek et al., 2008], many irrelevant candidates are still passed to the next step, which increase the potential number of false positives.

2.4.1.3 Segmentation and exhaustive search

Another approach is the combination of both techniques trying to leverage their strengths and address its drawbacks, but we have found only one example. In [Alonso et al., 2007], an initial selection of candidates is performed using segmentation with 3D information and, then, a second selection is performed using exhaustive search, but due to computational efficiency only around the center of those pre-selected candidates, i.e., bounded sliding-window.

2.4.1.4 Conclusions

Both approaches aim the generation or extraction of the initial object hypotheses (candidates to be a person) in the scene. So, they extract regions of interest from the image to be sent to the next processing module, avoiding as many background regions as possible. These techniques are of remarkable importance to reduce the number of candidates to be processed in the following stages, however, always keeping a balance between the number of candidates and the number of missing persons. Otherwise the number of false positive detections could be drastically increased or the subsequent modules will not be able to detect these missing persons, respectively.

The segmentation approach greatly facilitates the subsequent classification task, but it is affected by the inherited problems of the segmentation. In contrast, the exhaustive search approach provides a more robust candidate extraction, at the cost of increasing the subsequent classification task complexity and the global computational cost. The combination of both techniques can be a solution to merge their strengths and reduce their weaknesses.

2.4.2 Person model

As we have already commented, the verification or classification process applies a previously defined or trained person model to the objects candidates to be a person from an image or sequence and takes a final decision based on their similarity (see Figure 2.1). So, the definition of a proper person model is a critical task for the verification or classification process. There are two main discriminative information sources to characterize the people model: appearance and motion (see Figure 2.3). In any case, the model should be able to discriminate between people and any other object in the scene. Table 2.3 summarizes the different approaches from the state of the art according to the used person model information.

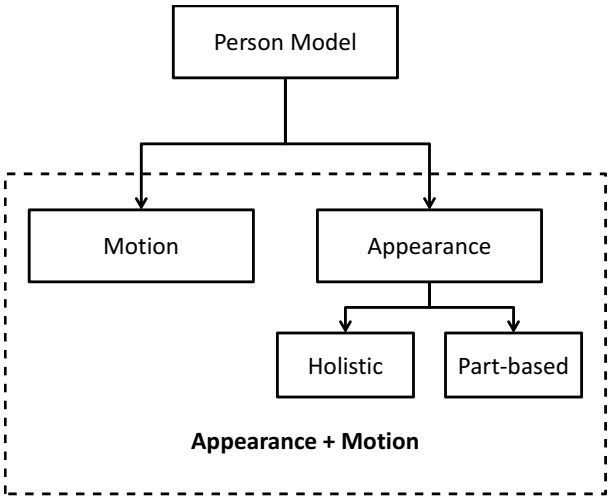


Fig. 2.3. People detection classification II.

2.4.2.1 Based on motion

Nowadays in the existing literature, most methods are only based on appearance information or they add robustness to the detection with motion information through tracking algorithms. However, human appearance varies due to environmental factors such as light conditions, clothing, contrast, etc, apart from the huge intrinsic people variability such as different heights, widths, poses, etc. For these reasons, there are some approaches which try to avoid these factors and to perform the detection using only motion information [Cutler and Davis, 2000; Sidenbladh, 2004].

Approach	Motion	Appearance	
		Holistic	Part-based
[Cutler and Davis, 2000]	Periodic motion	-	-
[Sidenbladh, 2004]	Flow patterns	-	-
[Xu and Fujimura, 2003; Zhao and Nevatia, 2004; Zhou and Hoang, 2005; Hussein et al., 2006; Gavrila and Munder, 2007; Koenig, 2007; Fernández-Carbajales et al., 2008; Kilambi et al., 2008]	-	Silhouette	-
[Viola and Jones, 2004]	-	Haar-like features	-
[Dalal and Triggs, 2005; Zhu et al., 2006; Zhang et al., 2007; Wojek et al., 2008]	-	HOG	-
[Leibe and Schiele, 2004; Seemann and Schiele, 2006; Leibe et al., 2008]	-	ISM	-
[Haritaoglu et al., 2000]	-	-	Silhouette
[Sprague and Luo, 2002; Harasse et al., 2006]	-	-	Color distribution
[Alonso et al., 2007]	-	-	Canny / Haar /.../ features
[Felzenszwalb et al., 2010]	-	-	HOG
[Wu and Nevatia, 2005]	-	-	Edgelets
[Andriluka et al., 2009]	-	-	ISM
[Viola et al., 2003; Cui et al., 2007]	Haar-like features multi-frame	Haar-like features	-
[Dalal and Triggs, 2006]	HOG multi-frame	HOG	-
[Giebel et al., 2004]	Tracking	Silhouette	-
[Okuma et al., 2004; Li et al., 2008; Ren, 2008]	Tracking	Haar-like features	
[Avidan, 2007; Breitenstein et al., 2010; Stalder et al., 2010; Yu et al., 2011]	Tracking	HOG	-
[Leibe et al., 2007]	Tracking	ISM	-
[Ess et al., 2009]	Tracking	HOG or ISM	HOG
[Wu and Nevatia, 2007]	Tracking	-	Edgelets
[Andriluka et al., 2008]	Tracking	-	ISM

Table 2.3: State of the art people detection classification II.

Within this classification, [Cutler and Davis, 2000] proposes an object classification system based on periodic motion analysis. The algorithm segments the motion, tracks objects in the foreground, aligns each object along time and finally computes the self-similarity between objects and how it evolves in time. Another approach based on motion information [Sidenbladh, 2004] proposes a people detection system based on detecting people motion patterns. For each object present in two consecutive images size normalization is performed and its flow pattern is calculated, that consists of dense optical horizontal and vertical flows.

In relation to people detection, methods based on motion usually obtain worse results than methods based on appearance, but they are independent of appearance variability. They do not support partial occlusions because in this case we could not extract motion patterns correctly. For these reasons they can only be considered either complementary information or essential in specific scenarios where methods based on appearance do not work (e.g., bad illumination, small objects).

2.4.2.2 Based on appearance

There are many approaches that use appearance information to define the person model. This is because appearance is more discriminant than motion. We classified the appearance models according to simplified human models or complex models. There are simple person models that define the person as a region or shape, i.e., holistic models [Viola et al., 2003; Xu and Fujimura, 2003; Giebel et al., 2004; Leibe and Schiele, 2004; Okuma et al., 2004; Viola and Jones, 2004; Zhao and Nevatia, 2004; Dalal and Triggs, 2005; Zhou and Hoang, 2005; Dalal and Triggs, 2006; Hussein et al., 2006; Seemann and Schiele, 2006; Zhu et al., 2006; Avidan, 2007; Cui et al., 2007; Gavrila and Munder, 2007; Koenig, 2007; Leibe et al., 2007; Zhang et al., 2007; Fernández-Carbajales et al., 2008; Kilambi et al., 2008; Leibe et al., 2008; Li et al., 2008; Ren, 2008; Wojek et al., 2008; Ess et al., 2009; Breitenstein et al., 2010; Stalder et al., 2010; Yu et al., 2011] and more complex models that define the person as combination of multiple regions or shapes, i.e., part-based models [Haritaoglu et al., 2000; Sprague and Luo, 2002; Wu and Nevatia, 2005; Harasse et al., 2006; Alonso et al., 2007; Wu and Nevatia, 2007; Andriluka et al., 2008, 2009; Ess et al., 2009; Felzenszwalb et al., 2010].

Within this classification (see Table 2.3), there are different chosen characteristics to define the people appearance, both holistic and part-based models. There are some approaches that extract the object silhouette and classify the object according to their similarity with reference people silhouettes or certain standards that the silhouette must meet. Some approaches make use of the color distribution in a person (where the skin color is essential) to determine if the object is a person or not. But the most popular approaches are those that define the people appearance according to their characteristic edge information using some kind of shape descriptor: Haar-like features [Viola et al., 2003; Viola and Jones, 2004; Okuma et al., 2004; Alonso et al., 2007; Cui

et al., 2007; Li et al., 2008; Ren, 2008], HOG (Histogram of Oriented Gradients) [Dalal and Triggs, 2005, 2006; Zhu et al., 2006; Avidan, 2007; Zhang et al., 2007; Wojek et al., 2008; Ess et al., 2009; Breitenstein et al., 2010; Felzenszwalb et al., 2010; Stalder et al., 2010; Yu et al., 2011], Edgelets [Wu and Nevatia, 2005, 2007] or ISM (Implicit Shape Model) [Leibe and Schiele, 2004; Seemann and Schiele, 2006; Leibe et al., 2007; Andriluka et al., 2008; Leibe et al., 2008; Andriluka et al., 2009; Ess et al., 2009].

Generally, those detectors based on a simplified or holistic person model have lower complexity, but do not support partial occlusions or pose variations. If you cannot see the whole region or shape, the model does not work properly. On the other hand, those detectors based on a more complex or part-based person model usually have higher complexity, but they support partial occlusions and pose variations.

2.4.2.3 Based on appearance and motion

Although the vast majority of approaches are mainly based on appearance information, there are some approaches that combine appearance and motion information in order to improve the detection results. Some authors combine appearance and motion expanding previous detectors based on appearance to more than one frame [Viola et al., 2003; Dalal and Triggs, 2006; Cui et al., 2007]; in this way they are able to introduce easily motion information in the person model and add robustness to the detector. Lately, the most popular approaches (detection-by-tracking approaches) are those that combine detection and tracking in order to improve the detection results [Giebel et al., 2004; Okuma et al., 2004; Avidan, 2007; Leibe et al., 2007; Wu and Nevatia, 2007; Andriluka et al., 2008; Li et al., 2008; Ren, 2008; Ess et al., 2009; Breitenstein et al., 2010; Stalder et al., 2010; Yu et al., 2011]. In this case, the motion information is not implicitly part of the person model, but it is still useful in order to filter or extrapolate detections over time.

2.4.2.4 Conclusions

As we have already commented, there are few approaches based only on motion information. Their main advantages are that they are independent of appearance variability and usually have low complexity. However they usually have worse results and they do not support partial occlusions.

The methods based on holistic person models (only a region or shape) usually have lower complexity, but they do not support partial occlusions neither pose variations. However, the methods based on part-based people models usually have higher complexity, but they support partial occlusions and pose variations. Another advantage is that they made the final decision by combining multiple evidences, so they are usually more reliable than methods based on holistic human models. For these reasons, they usually have better results.

Motion information can add robustness to the appearance model without adding too much complexity to the detection or even can be essential in specific scenarios where methods based on appearance do not work (e.g., tracking information could be very discriminant in complex scenarios which usually include multiple persons, multiple occlusions and background variability).

2.5 Summary and conclusions

During this chapter the different processing tasks that imply the automatic people detection have been analyzed. Then, a complete classification of the people detection approaches from the state of the art has been made regardless of their subsequent video surveillance application. Each classification includes a brief discussion about advantages and disadvantages of different approaches to solve the people detection problem in video sequences. This section sums up some conclusions extracted from the study performed.

As already explained in section 2.3, the people detection task consists mostly of, firstly, the design and training of a person model based on characteristic parameters (motion, dimensions, silhouette, etc); and, secondly, the adjustment of this model to the candidate objects in the scene. Thus, the critical tasks in any people detection algorithm are the generation or extraction of the initial object hypotheses to be people from the scene and the person model used to classify those initial object hypotheses.

The object detection approach has a great influence on the final people detection results. Firstly, every object not extracted during this stage cannot be classified as person. And secondly, a poor initial object extraction makes it more difficult the later classification. Segmentation is a simple and powerful object extraction technique, but with all their difficulties and limitations in complex environments. In contrast, the exhaustive search is more robust to rotation, scale and pose changes even in complex environments, but has the complexity of adding many false examples to the classification task, in addition to a higher computational cost.

The chosen person model to classify initial objects candidates to be person determines the robustness of the algorithm to person variations and occlusions. Simple models based only on motion or holistic appearance models are less robust to people variations and occlusions, whilst more complex part-based models add complexity to the algorithm, but they are much more robust to people variations and occlusions. Finally, the adequate combination of appearance and motion can improve the detection results.

In this chapter an overview of the state of the art in automatic people detection in video sequences has been presented. In Part II chapters 4, 5, 6, 7 and 8, the state of the art will be extended according to specific aspects defined for each chapter.

Chapter 3

People detection benchmarking framework

3.1 Introduction[1]

The ability to detect people in video is the key to a number of multiple applications, not only in video surveillance, but also in different areas like robotics, video games, intelligent vehicles, etc. Due to the rise in popularity of these applications over the last years, people detection has gradually experienced a great development. In parallel, interest on reliable strategies to assess the quality of people detection has also grown. Nowadays there are several public datasets that try to evaluate the performance of people detection algorithms. These datasets and a performance evaluation methodology are necessary to fairly evaluate algorithms under different conditions and to compare new algorithms with existing ones.

In this chapter, we describe the proposed people detection experimental setup. We describe the state of the art and the developed people detection experimental corpus, named PDds (Person Detection dataset), in section 3.2. Then, the people detection performance evaluation methodology is described in section 3.3. Finally, section 3.4 summarizes the chapter with some conclusions.

[1] This chapter is based on the publications *"A. García-Martín, J. M. Martínez, J. Bescós. A corpus for benchmarking of people detection algorithms. Pattern Recognition Letters, 33 (2): pp. 152-156, January 2012"* and *"J. C. SanMiguel, A. García-Martín, J. M. Martinez. Performance evaluation in video-surveillance systems: the EventVideo project evaluation protocols. Intelligent Multimedia Surveillance: Current Trends and Research, Pradeep Atrey, Mohan Kankanhalli, Andrea Cavallaro (eds.), 2013, Springer (in press)"*

3.2 Experimental corpus

3.2.1 Related work

Most of the reported people detection datasets are just based on sets of images [Papageorgiou and Poggio, 2000; Dalal and Triggs, 2005; Munder and Gavrila, 2006; Wojek et al., 2009]. There are also many video datasets in the video surveillance domain, but most of them do not include ground-truth annotations for people detection [Vezzani and Cucchiara, 2008]; they just include annotations for action recognition [AVSS; PETS; TRECVID]. A majority of the datasets including ground-truth annotations for people detection are designed only for specific surveillance applications: driver assistance systems [Wojek et al., 2009; Enzweiler and Gavrila, 2009; Dollár et al., 2009], people detection walking through a busy pedestrian zone [Ess et al., 2007], very specific scenarios [Andriluka et al., 2008] or even very general video surveillance systems [Nghiem et al., 2007].

Based on our experience in the field of people detection in video sequences, we describe a set of videos and annotations designed specifically for the people detection task. We have analyzed the critical factors that influence the detection and generated a corpus (dataset and associated ground-truth) in which they are specifically considered. Table 3.1 provides a detailed comparison of existing public people detection datasets.

As opposed to people detection datasets based on images, the availability of sequences of images inherent to a video dataset allows to consider motion information and to evaluate tracking based approaches. Additionally, according to the study and identification of critical factors affecting people detection techniques, we have designed a dataset that includes different background and people classification complexity levels (low, medium and high). The described dataset mainly excels other datasets in the amount of sequences (90 videos) and their variability. It includes a great variability of scenarios (outdoor and indoor surveillance scenes with different background complexities; textural, lighting changes, multimodal, etc) and a great variability of people appearance and interactions (scenes with one or multiple persons, pose changes, scale variations, people wearing different clothes, people carrying different objects and people with multiple interactions with objects and/or persons).

3.2.2 Proposed corpus

In the rest of this section, we describe the proposed people detection corpus (dataset and associated ground-truth). Firstly, we describe the design considerations necessary to achieve a representative set of video-sequences from a people detection point of view. Then, the sequences definition and annotation procedure are discussed and, finally, some examples are provided.

Name	Content	Numbers		Ground-truth	Complexity[2]
		Images[1]	Videos		
MIT [Papageorgiou and Poggio, 2000]	Color images	924 pos	-	Cut-outs images	Low: F/B views
INRIA [Dalal and Triggs, 2005]	Color images	902 pos 1671 neg	-	Bounding box (PASCAL format[PASCAL])	Low: F/S/B views
DCI [Munder and Gavrila, 2006]	Gray-scale images	24000 pos 25000 neg	-	Cut-outs images	Low: F/S/B views
TUD-Brussels [Wojek et al., 2009]	Color pair-images	508 pos	-	Bounding box (non-standard format)	Medium: F/S/B views, occlusions and multiple scales
TUD-MotionPairs [Wojek et al., 2009]	Color pair-images	1310 pos	-	Bounding box (non-standard format)	Medium: F/S/B views, occlusions and multiple scales
TUD-Pedestrians [Andriluka et al., 2008]	Color images/videos	860 pos	2 videos (272 frames)	Bounding box (non-standard format)	Medium: F/S/B views and occlusions
DCII [Enzweiler and Gavrila, 2009]	Color images/videos	15560 pos 6744 neg	1 video (21791 frames)	Bounding box and 3D localization (non-standard format)	High: F/S/B views, occlusions, multiple scales and non-static camera
Caltech [Dollár et al., 2009]	Color videos	-	1 video (250000 frames)	Bounding box (vvb file format)	High: F/S/B views, occlusions, multiple scales and non-static camera
ETHZ [Ess et al., 2007]	Stereo-Color videos	-	4 videos (2293 frames)	Bounding box (non-standard format)	High: F/S/B views, occlusions, multiple scales and non-static camera
PDds[3] (see section 3.2.2)	Color videos	-	90 videos (28358 frames)	Bounding box (Viper xml format [ViPER])	Low,Medium and High: F/S/B views, occlusions, multiple scales, interactions, backgrounds and static or non-static camera

Table 3.1: Public people detection datasets. [1]Number of positive (pos) and negative (neg) examples. [2]Views: front (F), side (S) and back (B). [3]Person Detection dataset (PDds) `http://www-vpu.eps.uam.es/DS/PDds/`.

3.2.2.1 Ground-truth design: critical factors in people detection

In order to obtain meaningful evaluation results, a corpus should include a set of representative video sequences, ranging from low to high complexity situations. The term "complexity" will be used hereinafter to express the degree of difficulty for a particular people detection algorithm to yield accurate results.

As already explained in the previous chapter, the people detection task [Hu et al., 2004; Valera and Velastin, 2005] consists mostly of, firstly, the design and training of a person model based on characteristic parameters (motion [Cutler and Davis, 2000], dimensions [Kilambi et al., 2008], silhouette [Xu and Fujimura, 2003], etc) and, secondly, the adjustment of this model to the candidate objects in the scene. All candidates that adjust to the model will be detected or classified as person, whilst all the others will not. Therefore, people detection can be split up into the localization of initial object candidates in the scene (object detection) and their subsequent classification (verification). Starting from these ideas, global sequence complexity has been found to be strongly dependent on a series of specific properties of objects [Dollár et al., 2009], on background complexity [Tiburzi et al., 2008] and on some relationships among these elements [Wu and Nevatia, 2005]. These dependencies have been designated as "critical factors", emphasizing their influence on the detection results. Since specific settings for these factors can significantly increase (low complexity settings) or decrease (high complexity settings) detection accuracy, they seem a convenient mechanism to regulate sequence complexity.

Table 3.2 summarizes the critical factors concerning foreground and background that we have considered. We next describe them including a brief discussion on their influence on the overall sequence complexity.

Background						Classification					
Textural complexity			Variability			Appearance variability			People/Object interactions		
Not textured	Slightly textured	Textured	Lighting changes	View changes	Multimodal	Pose variations	Different clothes	Carry objects	Objects	People	Objects & People

Table 3.2: Critical factors in people detection.

3.2.2.2 Background complexity

We here define background complexity as the difficulty to detect in the scene the initial objects candidate to be persons, due to the presence of edges, multiple textures, lighting changes, reflections, shadows and any kind of background variation. The following critical factors have been identified:

Textural complexity. Scenarios including an important amount of textured areas can make more difficult the localization of initial object candidates. In fact, depending on the algorithm used, highly textured background areas can be easily detected incorrectly as objects. Consequently, low textured background areas correspond to lower complexity situations and vice versa.

Variability. This refers to the property of some backgrounds to undergo variations usually produced by external factors (light and point of view changes) or multimodal backgrounds (such as twinkling water, swaying trees or glowing flames). Static scenarios with fewer variations correspond with low complexity levels, while scenarios with multiple variations correspond with more challenging situations.

3.2.2.3 People classification complexity

We here define people classification complexity as the difficulty to verify the object candidates to be person in the scene. It is related to the number of objects, their velocity, partial occlusions, pose variations and interactions among different people and objects. We have grouped these elements into two fundamental critical factors:

Appearance variability. People appearance exhibits very high variability since they are non-rigid objects, they can change pose, they can also wear different clothes or carry different objects, and they have a considerable range of sizes and shapes mainly due to the point of view and the relative situation with respect to the camera. People with limited appearance variability (no pose changes, no sizes variations, etc) entail low complexity levels, while the cases with high appearance variability entail a more complex classification.

People/Object interactions. People must be identified in real life scenarios, that is, they must be detected in the context of the environment surrounding them. People present interactions with objects and with other people. These interactions make more difficult their identification and classification. In order to identify all persons involved in these situations, it is necessary to deal with occlusions. Occlusions resulting from objects, other persons or visibility of the camera limits the visible appearance of the person occluded.

3.2.3 Description of the ground-truth

In the previous section, high, medium and low complexity settings for every critical factor have been identified. They have all been considered in the ground-truth design, thus making the resulting set of sequences specially useful to identify weak-points of a specific algorithm. We have grouped all the test sequences into different complexity categories and subcategories depending on these critical factors. A description of complexity levels for the associated content is shown in Table 3.3, whilst Figure 3.1 shows two examples of each category. The videos have been collected from several public datasets related with the people detection or object classification task [Vezzani and Cucchiara, 2008; Tiburzi et al., 2008], AVSS 2007 dataset ([AVSS]), PETS 2006 dataset ([PETS]) and TRECVID 2008 dataset ([TRECVID]).

Overall, sequences include both non-rigid (people, clothes, ropes, etc) and rigid objects (boxes, rucksacks, toys, etc) differing in size, motion (slow and fast displacements, rotations and chaotic motion) and textural appearance. These objects are involved in a number of interactions (intersecting and not intersecting trajectories, merging and splitting, partial and complete occlusions, etc) and in different contexts, like typical every-day situations (runners taking over each other, object being thrown, people dancing, etc) or surveillance video scenarios (office scenarios, subway platform, etc). Regarding the backgrounds, sequences include in-door and out-door scenarios. Additionally, different background complexities were considered by controlling the influence of homogeneous areas, external factors variations and multimodal motion.

3.2.4 Sequences annotation

In addition to video frames, a description of the detected people (frame number and bounding box) are also required in order to have the corpus ground-truth. Therefore, we have manually annotated 90 sequences (see Table 3.3). To carry out the annotation task, we have used the Viper tool [ViPER] that outputs XML files with the description (frame by frame people location, width and height). We have decided to use the Viper tool because it is one of the most popular ones in the research community, it is easy to manage, it has associated performance evaluation tools and it offers a variety of metrics for performing comparison between video metadata files.

In complex environments with multiple people and partial occlusions, it is often not obvious where to draw the line and decide whether a person should be annotated or not. In our set of sequences, people "occur" in every state of occlusion, from fully visible to just one single body part visible. We therefore decided to annotate all those cases where a human could clearly detect the person, without human reasoning. As a consequence, all people were annotated as a single entity (blob) covering the connected or disconnected visible parts of them whenever at least the head or most of the torso is visible.

Sequence	Category	Subcategory	Background		Classification	
			Textural complexity	Variability	Appearance variability	People/Object interactions
1-4	C1	C1-a	Low	Low	Low	Low
5-6	C1	C1-b	Low	Medium	Low	Low
7-8	C2	C2-a	Low	Low	Medium	Low
9-10	C2	C2-b	Low	Low	Medium	Medium
11-12	C2	C2-c	Low	Medium	Low	Medium
13	C3	C3-a	Medium	Medium	Medium	Low
14-16	C3	C3-b	Medium	Medium	Medium	Medium
17-18	C4	C4-a	Low	Low	Medium	High
19-20	C4	C4-b	Low	Low	High	Medium
21	C4	C4-c	Low	Low	High	High
22-24	C5	C5-a	Medium	High	Medium	High
25	C5	C5-b	Medium	High	High	Medium
26	C5	C5-c	High	High	Medium	High
27-33	C5	C5-d	High	High	High	Low
34-65	C5	C5-e	High	High	High	Medium
66-90	C5	C5-f	High	High	High	High

Table 3.3: Critical factors on experimental dataset.

3.2.5 Examples

Figure 3.1 shows some example frames from several sequences of the corpus including annotated blobs, just to offer an idea of sequences appearance and their corresponding annotations. The complete set of sequences along with their description, associated category and the annotation ground-truth can be downloaded[2] (Figure 3.2 shows a screenshot of the web site).

3.3 Performance evaluation methodology

This section describes the experimental setup or evaluation methodology. In order to define a proper evaluation methodology, it is necessary to define the chosen evaluation video corpus (or dataset) and the chosen evaluation metrics.

[2] `http://www-vpu.eps.uam.es/DS/PDds/` It is freely available for research purposes (after completing a license agreement form).

Fig. 3.1. Experimental dataset examples. Every example shows three random frames from a sequence.

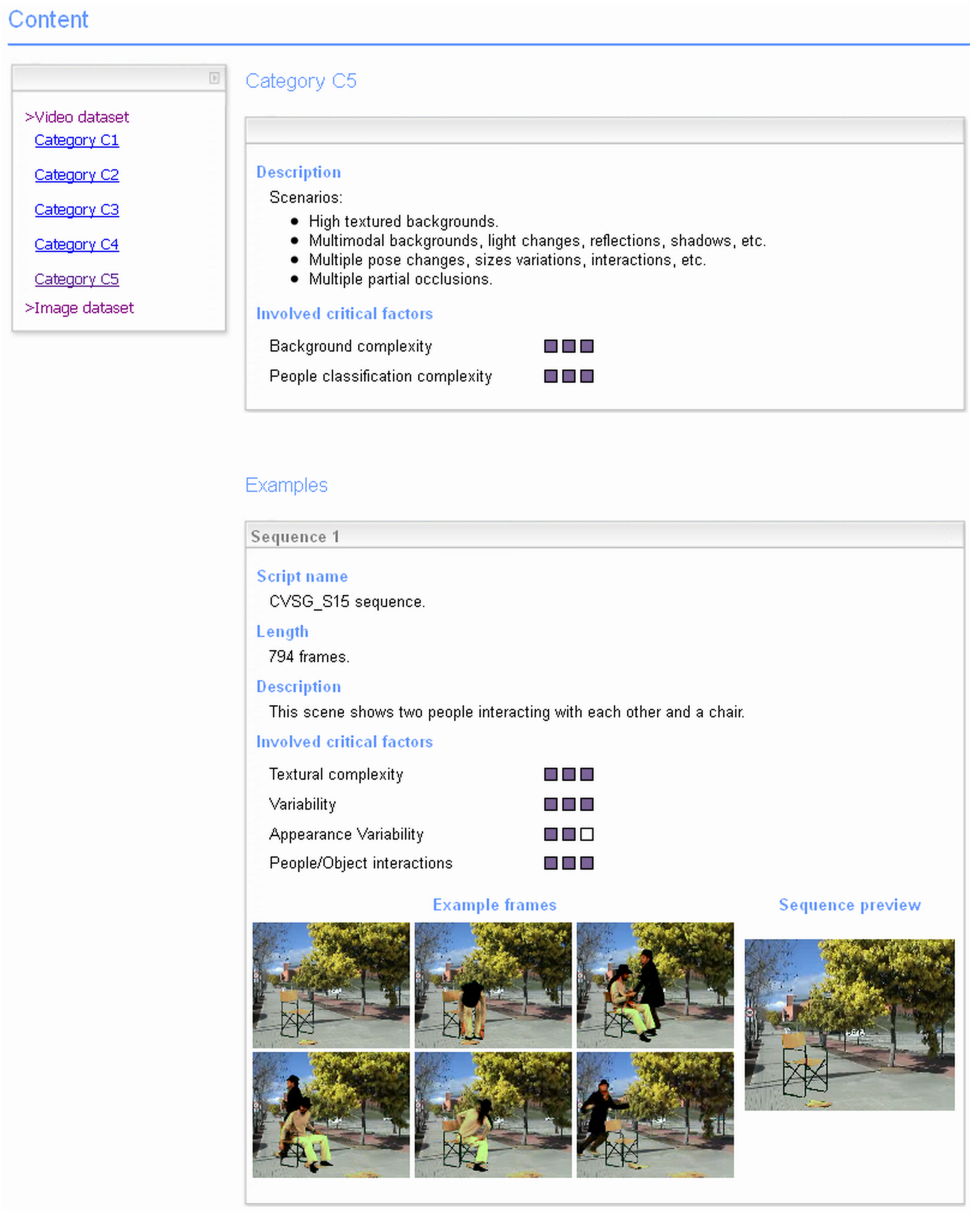


Fig. 3.2. Screenshot of the experimental dataset public web.

3.3.1 Evaluation dataset

The experimental corpus PDds (see section 3.2) has been divided in two evaluation datasets. The first dataset, named A, has been selected to evaluate the different approaches at every complexity level; it includes the first 29 sequences from our experimental corpus. These sequences include the five different complexity categories depending on the defined people detection critical factors. As already commented, the experimental dataset includes both non-rigid and rigid people and objects differing in size, motion and textural appearance. These people and objects are involved in a number of interactions and in different contexts, like typical every-day situations or surveillance video scenarios. Regarding the backgrounds, it includes in-door and out-door scenarios with different background complexities.

The second dataset, named B, has been selected to evaluate more thoroughly the category with the highest complexity, i.e., category C5. It includes the following 61 sequences from our experimental corpus. The sequences have been extracted from the TRECVID 2008 dataset [TRECVID], namely, the ones for the surveillance TRECVID event detection task recorded at London Gatwick International Airport. This dataset contains highly crowded scenes, severely cluttered background, people at different scales and people completely static along the whole sequences. Due to the small size of the objects at the top of the image, during the annotation of sequences, the top 15% of the images has been discarded. Figure 3.3 shows some examples of final annotations.

A summary of the complexity levels of both evaluation datasets is shown in Table 3.4.

Category	#Sequences		Complexity	
	Dataset A	Dataset B	Classification	Background
C1	6	0	Low	Low
C2	6	0	Medium	Low
C3	4	0	Medium	Medium
C4	5	0	High	Low
C5	8	61	High	High

Table 3.4: Sequences categorization evaluation datasets.

3.3.2 Evaluation metrics

In order to evaluate different people detection approaches, we need to quantify the different performance results. In the state of the art, performance can be evaluated at two levels: sequence sub-unit (frame, window, etc) or global sequence. Sub-unit performance is usually measured in terms of Detection Error Tradeoff (DET) [Dalal and Triggs, 2005; Dollár et al., 2012] or Receiver Operating Characteristics (ROC) [Munder and Gavrila, 2006; Enzweiler and Gavrila,

Fig. 3.3. Ground-truth examples of evaluation dataset B. 15% of the top (in red) has not been considered for evaluation purposes.

2009] curves. Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [Andriluka et al., 2008; Leibe et al., 2008; Wojek et al., 2009]. The first level gives us information about the classification stage, while the second one provides overall system performance information. In order to evaluate a video surveillance system, it is more interesting to compare the overall performance. In both cases the detectors output is a confidence score for each person detection, where larger values indicate higher confidence. Both evaluation methods compute progressively the respective parameters such as the number of false positives, Recall rate or Precision rate from the lowest possible score to the highest possible score. Each score threshold iteration provides a point on the curve.

ROC curves represent the fraction of true positives out of the positives (True Positive Rate -TPR-, Recall or Sensitivity) vs. the fraction of false positives out of the negatives (False Positive Rate -FPR- or 1-Specificity). We aim to evaluate and compare the overall performance of different detection systems, so we have chosen the second evaluation method. For each value of the detection confidence, Precision-Recall curves compute Precision and Recall as follows:

$$Precision = \frac{\#TruePositivePeopleDetections}{\#TruePositivePeopleDetections + \#FalsePositivePeopleDetections} \tag{3.1}$$

$$Recall = \frac{\#TruePositivePeopleDetections}{\#TruePositivePeopleDetections + \#FalseNegativePeopleDetections} \tag{3.2}$$

In order to evaluate not only the yes/no detection decision, but also the precise persons locations and extents, we use three evaluation criteria, defined by [Leibe et al., 2005], that allow to compare hypotheses at different scales: relative distance, cover and overlap. The relative distance dr measures the distance between the bounding box centers in relation to the size of the annotated bounding box. Cover and overlap measure how much of the annotated bounding box is covered by the detection hypothesis and vice versa (see Figure 3.4). A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%. Only one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive.

The integrated Average Precision (AP) is generally used to summarize the overall performance, represented geometrically as the area under the PR curve (AUC-PR); in order to express more clearly the results, we have chosen the representation Recall vs. 1-Precision (see Figure 3.5). In order to approximate correctly the area, we use the approximation described by [Davis and Goadrich, 2006]. In addition, focusing on the people detection evaluation in video surveillance systems, we want also to evaluate the detector at the operating point, i.e., at the predefined optimal decision threshold for each algorithm. Thus we can compare the final operational per-

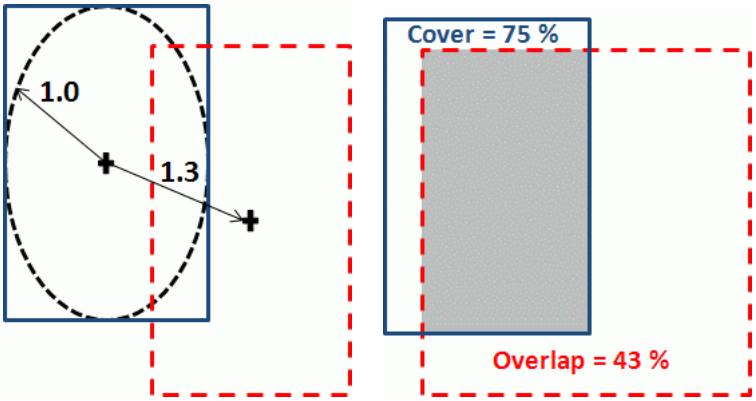
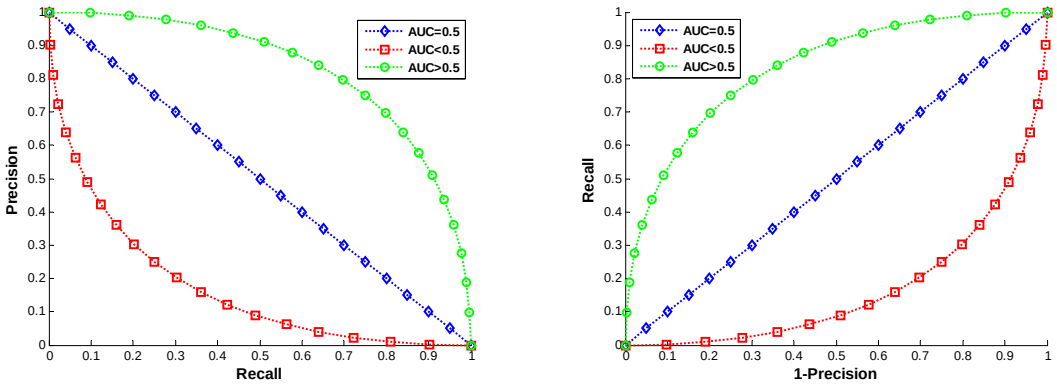


Fig. 3.4. Evaluation criteria for comparing bounding boxes [Leibe et al., 2005]: (left) relative distance; (right) cover and overlap.



(a) (b)

Fig. 3.5. Precision-Recall curves and area under the curve. Equivalent representations: (a) Precision vs. Recall representation and (b) Recall vs. 1-Precision representation.

formance and not just its overall performance.

3.4 Summary and conclusions

This chapter compiles the motivations and considerations applied to the generation of a corpus (dataset and associated ground-truth) and the definition of a performance evaluation methodology, for the evaluation of people detection algorithms in video sequences.

Both the wide range of considered critical factors and the development of an accurate ground-truth for the presented corpus, makes it especially suitable for tuning the algorithms, results evaluation and comparison. A more complete people detection corpus in surveillance scenarios than the ones available in the state of the art has been developed, providing a complete framework

for the evaluation of people detection algorithms under different complexity conditions.

Based on the state of the art, a people detection evaluation methodology has been defined with a particular interest in assessing the overall detection system performance instead of just the binary classifier performance (person/no person). Thus, we have chosen the Precision-Recall metrics and some additional evaluation criteria (relative distance, cover and overlap), all widely used to measure the accuracy of the people detection and localization.

Part II

People detection approaches

Chapter 4

Real time people detection based on appearance information

4.1 Introduction[1]

As already mentioned, the complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, poses, movements, points of views and interactions between different people and objects. Currently, many different systems exist which try to solve this problem. The state of the art in people detection and tracking includes several successful solutions working in specific and constrained scenarios. Most of them obtain good detection results, but do not operate in real time. In contrast, the systems operating in real time usually get worse results. The main objective of this chapter is to present a robust people detector that is capable of operating in real time. The work presented in this chapter is inspired by a well-established non-real time solution in the field [Wu and Nevatia, 2005], on which we introduce some useful modifications to operate in real time and add robustness to the detection.

In this chapter, we will firstly make a brief introduction of the literature related to real or non-real time people detection approaches in section 4.2. Then, the proposed real time people detection approach is described in section 4.3. After that, section 4.4 describes the experimental results. Finally, section 4.5 summarizes the chapter with some conclusions.

[1]This chapter is based on the publication *"A. García-Martín, J. M. Martínez. Robust Real Time Moving People Detection in Surveillance Scenarios. In Proc. of the IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 241-247, 2010"*

4.2 Related work

Focusing on the idea of a real video surveillance system, people detection algorithms can be classified into two main families depending on whether they work in real time or not, splitting the problem, and even the approach used in each case, in two systems clearly differentiated. On the one hand, systems that operate in real time usually get initial candidates location using image segmentation. Some approaches employ background subtraction [Zhao and Nevatia, 2004; Zhou and Hoang, 2005], whilst other approaches use stereo-vision or 3D information [Gavrila and Munder, 2007]. Besides, due to computational constraints, these approaches usually employ simplified person models (ellipse, human shape templates, etc). On the other side, the systems that do not operate in real time [Andriluka et al., 2009; Dalal and Triggs, 2005; Leibe and Schiele, 2004; Seemann and Schiele, 2006; Viola et al., 2003; Wu and Nevatia, 2005, 2007; Felzenszwalb et al., 2010] get these initial candidates locations scanning the complete image at various scales and rotations (exhaustive search); in this case, person models must be complex to classify correctly many negative examples. The scanning and use of more complex models improve the detection rate, but the computational costs are too high to allow for real time processing. Although some proposals have studied this problem [Zhu et al., 2006; Zhang et al., 2007; Wojek et al., 2008], many irrelevant candidates are still passed to the classification step, which increases the potential number of false positives.

There are some approaches that combine both techniques trying to leverage their strengths and address their drawbacks [Alonso et al., 2007]: an initial selection of candidates is performed using segmentation with 3D information and, then, a second selection is performed using exhaustive search, but due to computational efficiency only around the center of those pre-selected candidates, i.e., bounded sliding-window. For this reason, the algorithm still has a strong dependence with the previous segmentation, especially in partial object detections or overlapping objects.

As in the previous case [Alonso et al., 2007], we propose to combine segmentation and exhaustive search in order to achieve robustness and real time operation. Firstly, the initial objects candidates to be person are extracted using background subtraction and, then, those selected candidates are processed with an exhaustive search, in this case with a full exhaustive search over the selected candidates. In this way we maintain the positive aspects of the segmentation: initial candidates, easy rejection of irrelevant areas of the image and simpler person model focused on classification; and also the positive aspects of exhaustive search: robustness to scale and pose variations. The segmentation reduces the exhaustive search critical factor, eliminating easily most of the false negative examples. We still depend on segmentation, but in this case it is less critical because we use full exhaustive search over the already detected objects, being still robust to partial detections and overlapping objects.

4.3 Real time moving people detection

4.3.1 System overview

As already described in chapter 2, the basic architecture of any people detector includes an image or video input, the object detection task, the designed and trained (if training is required) person model, the verification/classification task and the final detection decision (see Figure 2.1). In order to evaluate our proposal in a "canonical" automated video analysis system for people detection (especially in terms of computational cost), the object tracking task has also been added (see Figure 4.1).

Input

The system is based on color frames extracted from static and mono camera video sequences.

Object detection

The selected object detection technique is background subtraction. It is a commonly used technique for motion detection and segmentation. Motion detection aims at segmenting regions corresponding to moving objects from the rest of the image. The consecutive stages depend on the background accuracy obtained, that is, the rest of stages have a strong dependency with the results obtained in this process: a bad background model could cause false object detections, missing objects or partial object detections. In our system, foreground extraction is based on [Cavallaro and Ebrahimi, 2001]. After segmentation, morphological operations are typically applied to reduce the noise of the resulting image mask and improve the object extraction [Valera and Velastin, 2005]. In our system, after object detection, a connected component analysis is applied [Dillencourt et al., 1992]. Only objects extracted in this stage are analyzed in following stages. Each object is defined with a blob (localization and dimensions).

Person model

The chosen person model corresponds to a simplified version of the person model defined by [Wu and Nevatia, 2005]. The details of this module in our system are described with more detail in section 4.3.2.

Object tracking.

After motion detection and object extraction, surveillance systems generally track moving objects. The aim of an object tracker is to generate the trajectory of an object over time by locating its position in every frame of the video sequence where it appears. In our system, a simple tracking algorithm based on the Kalman filter [Broida and Chellappa, 1986] is used and generates the trajectories of the blobs between consecutive frames using color information, the dimensions of the blob (width and height) and the position of the blob (centroid). The color information is the Hue-channel color histogram calculated on all points inside of the object mask.

Verification/Classification

A cascade Adaboost algorithm [Viola and Jones, 2001] (Gentle variation) is used to generate

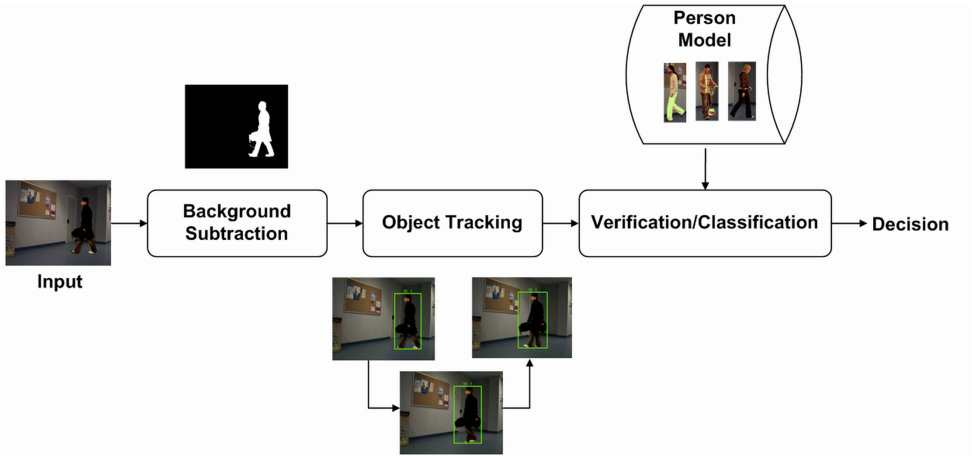


Fig. 4.1. Overall system architecture.

the final detection people confidence for each object candidate to be a person. The details of this module in our system are described with more detail in section 4.3.2.3.

Decision

For performance evaluation methodology purposes (see section 3.3), the output detection confidence is evaluated progressively in terms of Precision and Recall from the lowest possible score to the highest possible score PR curve, as well as with the AUC-PR to summarize the overall performance.

4.3.2 People detection approach

Our people detector is based on the algorithm proposed in [Wu and Nevatia, 2005], but proposing modifications in order to achieve real time performance in video surveillance scenarios.

4.3.2.1 Base person model

[Wu and Nevatia, 2005] proposes a method for human detection in crowded scenes, but working only with static images (frames). An individual human is modeled as an assembly of natural body parts. The main idea consists of identifying characteristic edges of each body part and generating four edge models of body parts (body, head, torso and legs). The image is scanned with four independent edge feature detectors previously trained. The training phase is performed using the Real Adaboost algorithm [Freund and Schapire, 1997] and a nested cascade structure [Huang et al., 2004]. Responses of each part detectors are combined to obtain a joint likelihood model that includes cases of multiple and possibly inter-occluded humans. This algorithm also supports changes in pose or camera point of view.

4.3.2.2 Proposed person model

The base algorithm is targeted to static images and scans the complete image; for these reasons person models must be complex in order to be able to classify correctly many negative examples. In addition, as at this phase computation time is not a main objective, the training phase is focused on reducing false positive rate (complex person models) what greatly increases the processing time. In order to get a faster algorithm, we propose not to scan the complete image and to simplify the person model. Firstly, instead of scanning the complete image, we only process moving objects detected in previous stages (see Figure 4.1). Secondly, the model of each body part is simplified (and consequently the final person model) what reduces the time needed during the detection process. The proposed modifications are the following: we use a ranking of the best edges of each body part and the training phase is not focused on reducing false positive, but also on getting good Precision results.

Edge shapes In this work, according to the size of the images (58x24 pixels) and the base method [Wu and Nevatia, 2005], the possible length (k) of one single edge is from 4 pixels to 12 pixels. The edge features we use consist of single shapes, including lines, 1/8 circles, 1/4 circles and 1/2 circles. We use 36 types of lines (four orientations: 0^o, 45^o, 90^o and 135^o; and 9 dimensions: 4-12 pixels; $4\,orientations \times 9\,dimensions = 36$). We generate arcs from 4 pixels to 12 pixels such that the perimeter of their circumference (P) follows:

Finally, we have a total of 775 edges (36 lines and 739 arcs). For example, when the size of the body image is 58×24, the overall number of possible edge features is 1.078.800 ($58 \times 24 \times 775 = 1.078.800$).

Learning part detectors For each edge feature, one weak classifier [Wu and Nevatia, 2005] is built. Then, the Adaboost algorithm [Freund and Schapire, 1997] is used to learn strong classifiers. The Adaboost algorithm has many variations such as Discrete Adaboost, Real Adaboost and Gentle Adaboost. Instead of using the Real Adaboost variation, Gentle Adaboost is chosen because it outperforms other variations as reported by [Lienhart et al., 2003]. In order to reduce computational cost and to identify the most characteristic edges of each body part, we make a top-100 edge ranking. We iteratively train in a bootstrap way the best classifier of each edge and select the best 100 associated edges. Finally, instead of using a complex nested structure focused on reducing the false positive rate, the cascade Adaboost algorithm [Viola and Jones, 2001] (Gentle variation) is used to learn each detector. This training phase is not only focused on reducing the false positive rate, but also on getting good Precision results.

4.3.2.3 Verification/Classification

Only objects detected after the previous stages (see Figure 4.1) are classified. Each corresponding image of each blob is normalized and, then, the four models of body parts (Gentle Adaboost cascade classifiers) are generated.

The classification process consists of evaluating the four models of body parts, providing four independent evidences. The final evidence about the analyzed blob being a person is obtained by averaging the evidences provided by the four body parts detectors.

The final people detection is less complex (by using a simplified person model and a smaller number of classifiers: those which belong to the top-100 edge ranking) and the completed system is faster (by not scanning the entire image), whilst maintaining good Precision results.

4.4 Experimental results

In this section, we describe the experiments carried out for testing the proposed people system over our video dataset and we compare the results of our approach, named Edge, with five other people detectors approaches from the state of the art: four non-real time approaches, HOG [Dalal and Triggs, 2005], ISM [Leibe et al., 2005], TUD [Andriluka et al., 2009] and DTDP [Felzenszwalb et al., 2010] detectors and one real time approach, Fusion [Fernández-Carbajales et al., 2008]. Our approach Edge, is based on [Wu and Nevatia, 2005], the authors themselves show in [Wu and Nevatia, 2007] similar results than HOG in terms of classification accuracy. On the other hand, TUD detector outperforms HOG detector, previous authors detector partISM [Andriluka et al., 2008] and ISM variations (4D-ISM [Seemann and Schiele, 2006] and standard ISM [Leibe et al., 2005]), in terms of classification accuracy. Finally, DTDP is a part-based adaptation of the original HOG. There a brief description of the different people detection approaches used from the state of the art in appendix A.

Experimental results include an evaluation of people detection rates and computational cost. The system has been implemented in C++, using the OpenCV image processing library[2]. The tests have been performed on a Pentium IV with a CPU frequency of 2.4 GHz and 3GB RAM. The Fusion results have been obtained with the original code, the HOG results have been obtained using the available binaries[3], the ISM results have been obtained using the available code and binaries[4], the TUD results have been obtained using the available code[5] and the DTDP results have been obtained using the available code[6].

[2] `http://sourceforge.net/projects/opencv/`
[3] `http://pascal.inrialpes.fr/soft/olt/`
[4] `http://www.vision.ee.ethz.ch/~bleibe/index.html`
[5] `http://www.d2.mpi-inf.mpg.de/andriluka_cvpr09`
[6] `http://www.cs.brown.edu/~pff/latent/`

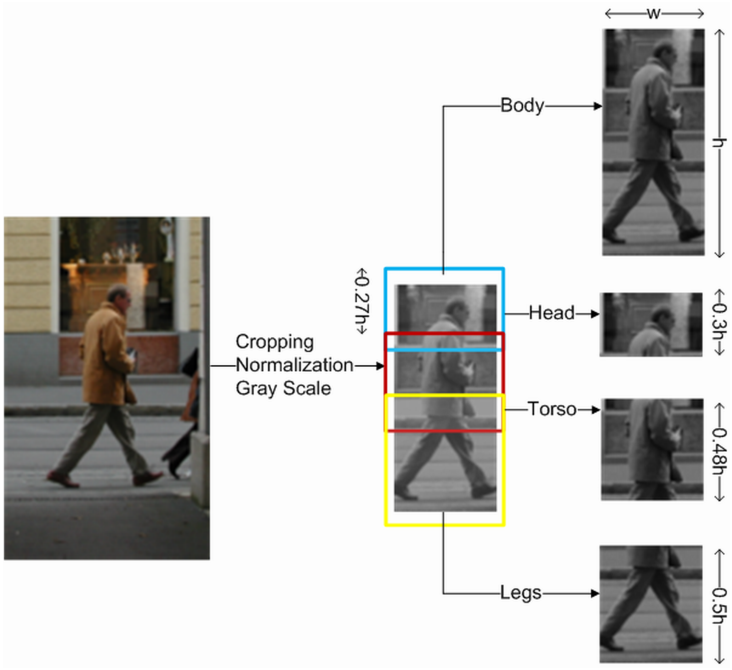


Fig. 4.2. Body part segmentation.

4.4.1 Experimental setup

4.4.1.1 Image training dataset

The proposed algorithm consists of four models of edge body parts. Each model has to be trained with an image collection with people and non-people examples and, therefore, we need a complete image dataset with positive and negative examples.

Negative images have been chosen from the LabelMe dataset [Russell et al., 2008]. Each image has been cropped in small pieces in order to obtain a huge number of different negative images. Positive images have been chosen from the INRIA dataset [Dalal and Triggs, 2005]. Person body blobs have been extracted, normalized (58x24 pixels and gray scaled) and segmented in body parts (see Figure 4.2) according to the base algorithm [Wu and Nevatia, 2005]. Finally, our image dataset stores 3.542 positive images (already extracted, normalized and segmented) and more than 40.000 negative images.

4.4.1.2 Video evaluation datasets

To evaluate our proposed people detection approach and compare with the state of the art approaches, it has been evaluated in both evaluation datasets (A and B) described in the performance evaluation methodology (see section 3.3). The dataset A allows us to evaluate the different approaches at every complexity level (C1,..,C5), while the B dataset allows us to evaluate more

thoroughly the highest complexity category (C5).

4.4.2 People detection results

Despite the fact that all algorithms performance depends on the hit rate, or confidence level of the decision, we only classify objects detected in previous stages (see Figure 4.1) as person or non-person. Consequently, the maximum and minimum Recall and Precision will be limited by previous stages. One of the previously mentioned approaches, Fusion, is based on the same scheme and also performs in real time. Moreover, the non-real time approaches, HOG, ISM, TUD and DTDP, the maximum and minimum Recall and Precision will be limited by the image scanning process.

4.4.2.1 Evaluation dataset A

Firstly, we evaluate and compare all approaches at every complexity level using the evaluation dataset A. Figure 4.3 shows the detection performance in terms of Recall vs. (1-Precision) curves (see section 3.3.2) on some examples of different complexity categories included within the used video dataset A. At low levels of classification and background complexity C1 -Figure 4.3 (a) and (b)-, our method, Edge, outperforms or obtains the same results than the other approaches.

At intermediate complexity categories C2, C3 and C4, our proposal is clearly superior to non-real time systems. In particular, we can see how in sequences with many partial occlusions and pose variations -Figure 4.3 (e), (f), (g) and (h)-, the non-real time systems performance is significantly reduced. While in our system even if some individual parts detectors may have poor results (partial occlusion and pose variations), the combined detector maintains high detection rates.

At high levels of classification and background complexity C5, the global performance of our system is reduced, mainly, due to the high background complexity. The first example -Figure 4.3 (i)- shows a scene with lighting changes, reflections and shadows; while the second example -Figure 4.3 (j)- shows a scene with lighting changes, shadows and a multimodal background (moving branches). However, the results are still better, or slightly better, than the best non-real time performance (ISM or DTDP respectively).

Non-real time approaches are robust due to the exhaustive person search carried out and their complex person models. However, in some cases they show an unreliable performance because of the high number of false positive examples that appear during exhaustive search. The previously explained problem affects to every non-real time approach: for HOG algorithm in Figure 4.3 (e), (i) and (j), for ISM and DTDP algorithms in Figure 4.3 (f) and for TUD algorithm in Figure 4.3 (e), (f) and (i). The real time approach, Fusion, also shows an unreliable performance. Its usage of a highly simplified person model achieves fast people detection, but not quite robust; as result, this approach presents an irregular behavior in all categories.

In our system, even though we use a simplified person model in order to work in real time, our performance is, in general, equal or superior to other approaches in all categories of the evaluation dataset A.

Table 4.1 shows the results in terms of AUC-PR (see section 3.3.2) for each complexity category of dataset A. Again, at complexity categories C1, C2, C3 and C4, our proposal is clearly superior to other approaches. All algorithms perform worse at higher complexity categories. However, it is observed that all approaches obtain generally worse results at category C3 than at category C4, due to the great influence of the background complexity in category C3 and, thus, the generation or extraction of the initial object hypotheses or candidates to be a person in the scene is more difficult. On the other side, the complexity of the category C4 lies on the classification of those initial candidates.

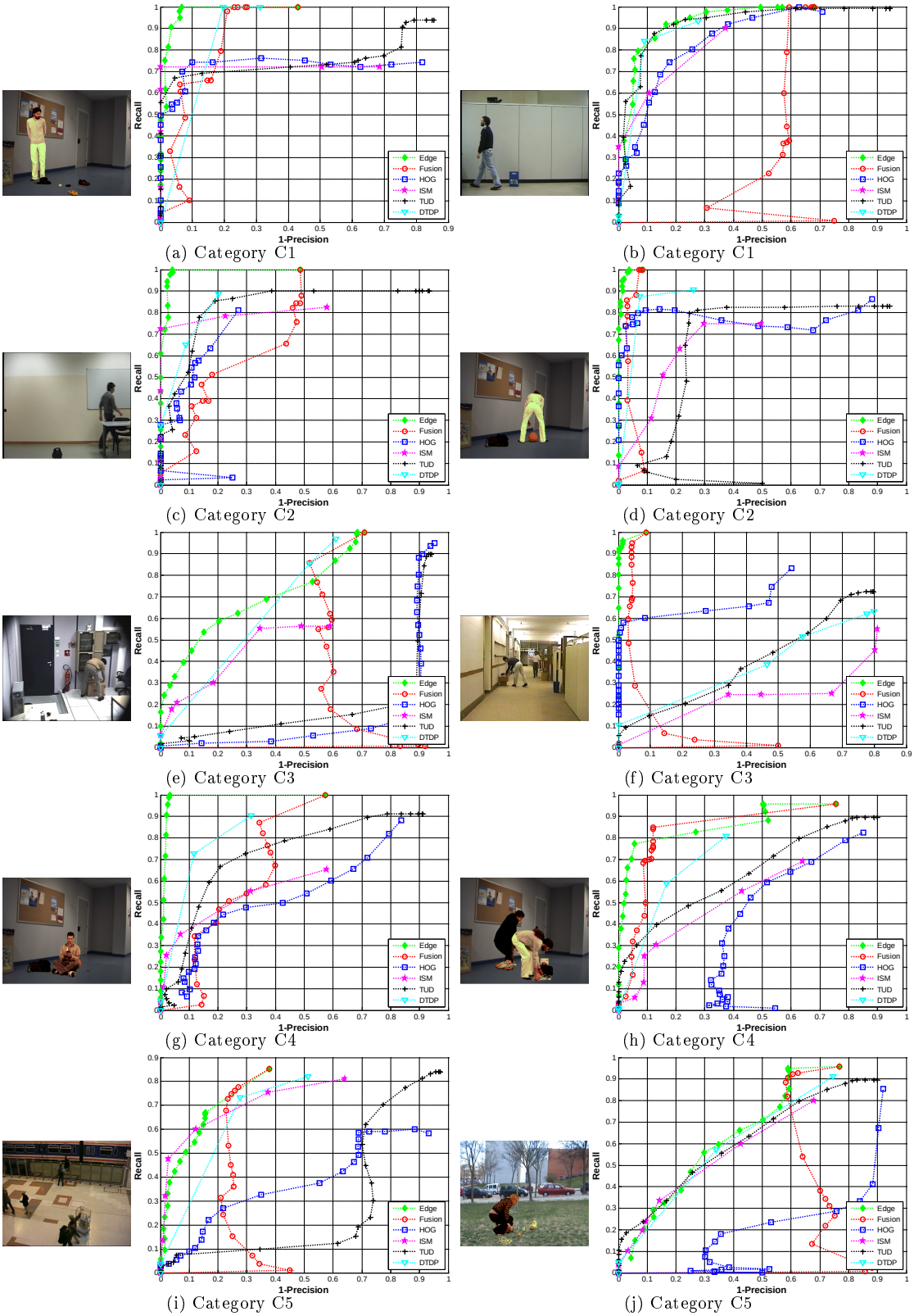
The Fusion approach gets the worst results. The use of segmentation simplifies the classification stage, allowing the approach to reach high recall results, but the use of such a simple person model together with some segmentation problems (under and over segmentation) reduce the global precision rate. The Edge approach gets good results in all complexity categories and similar to the other approaches not based on segmentation. It is due to the use of a more complex person model and the combination of segmentation and exhaustive search. Despite the fact that the combination of segmentation and exhaustive search reduces the segmentation problems, these problems are magnified in complex background scenarios (C3-C5) where it is quite difficult to obtain a reliable segmentation.

The exhaustive search approaches are more robust to scale and pose variations and, therefore, more reliable in complex environments than those based on segmentation. Even so, the background complexity still have a negative impact in the results (C3). Moreover, unlike the previous case, the classification stage is not simplified; it is even more complex because the approach must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. The HOG and TUD approaches show similar results in all complexity categories, but the ISM and DTDP get better results. The ISM is an holistic approach, but with a great flexible person model based on spatial feature probability distribution and the DTDP is a body part-based variation of the HOG approach.

4.4.2.2 Evaluation dataset B

In this section, we evaluate more thoroughly the highest complexity category (C5) using the dataset B. Table 4.2 shows the results in terms of AUC-PR of dataset B. Due to the greater complexity of the sequences extracted from TRECVID (the content set contains challenging scenarios, crowds and a wide range of scales), the results are worse than those obtained in the dataset A.

In this case, our approach obtains worse results than the non-real time approaches from the



(a) Category C1 (b) Category C1

(c) Category C2 (d) Category C2

(e) Category C3 (f) Category C3

(g) Category C4 (h) Category C4

(i) Category C5 (j) Category C5

Fig. 4.3. Examples of the sequences categories and people detection results on dataset A.

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C1	0.98	0.78	0.92	0.95	0.93	0.96	0.92
C2	0.93	0.81	0.86	0.91	0.88	0.92	0.89
C3	0.85	0.60	0.74	0.80	0.75	0.81	0.76
C4	0.89	0.69	0.82	0.84	0.84	0.86	0.82
C5	0.70	0.48	0.71	0.71	0.67	0.74	0.67
Total	0.87	0.67	0.81	0.84	0.81	0.86	0.81

Table 4.1: Area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A.

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.59	0.44	0.66	0.69	0.56	0.68	0.60

Table 4.2: Area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B.

state of the art. As already commented, the main problem of our approach is the difficulty of making a reliable segmentation (foreground and background) in complex scenarios. However, the sequences extracted from TRECVID present an additional difficulty to our approach: the sequences include people completely static along the whole sequences. Our approach extracts the objects candidates to be a person using motion information (background subtraction), being not able to extract static people or objects, what reduces the Recall rate and, therefore, the overall performance.

The results also show that the approaches based on exhaustive search also get worse results than with dataset A. However, except the TUD approach, they are more stable in more complex scenarios because they are more robust to scale and pose variations and more robust to the background complexity.

4.4.3 Computational cost

In this section, the computational cost, measured as the output processing rate in frames per second (fps), generated by our approach will be compared with two different approaches. In first place, the previously mentioned real time approach and, secondly, the non-real time approach called HOG. ISM, TUD and DTDP algorithms were also previously mentioned, however, due to their high computational cost they will not be considered in the comparison[7]. The above described dataset includes different video resolutions; the results obtained with 352x288 images are summarized in Table 4.3.

The computational cost of real time approaches depend a lot on each sequence. It does not only depend on people detection and background complexity, but also on many other factors:

[7]NC: Not considered in the comparison due to its high computational cost.

object dimensions, number of tracked objects, etc. For this reason, we show a summary with the worst, best and average results obtained over our proposed dataset.

The results show clearly how our proposed detector, Edge, works in real time and even faster than the previously mentioned real time approach, *Fusion*. Both real time approaches computational costs depend on the different sequences. Nevertheless, the non-real time approach remains almost invariant to different sequences because of the exhaustive search carried out.

	Edge	Fusion	HOG	ISM	TUD	DTDP
Minimum	64.5	14.5	11.4	NC[7]	NC[7]	NC[7]
Average	71.6	32.6	11.5	NC[7]	NC[7]	NC[7]
Maximum	80.8	62.8	11.6	NC[7]	NC[7]	NC[7]

Table 4.3: Computational cost: average frames per second (fps).

4.5 Summary and conclusions

In this chapter, an approach that combines segmentation and exhaustive search in order to achieve robustness and real time operation is presented. A complete surveillance video system has been implemented to evaluate the proposed detection approach. Besides, in order to provide a good performance evaluation of the proposed framework, it has been evaluated over the PDds composed of several annotated surveillance sequences of different levels of complexity.

Experimental results over the proposed evaluation dataset A show that the proposed system performs considerably well at real time and even better than other non-real time approaches from the state of the art and that it is significantly more efficient and stable than others approaches from the state of the art. However, due to the background segmentation difficulty in complex scenarios, at high levels of complexity our proposal obtains similar results than the state of the art.

Experimental results over the proposed evaluation dataset B points out that our approach does not work properly in more complex and realistic scenarios. Our approach presents a strong dependence with the segmentation stage, so all the segmentation problems are inherited (under and over segmentation). Our combination of segmentation and exhaustive search reduces these problems, but these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation.

In the following chapters, we propose new algorithms that improve the state of the art in more complex and realistic environments, making use of approaches based on exhaustive search, since they are more robust in complex scenarios. These approaches will imply more restrictions in order to fulfill the objective of real time performance.

Chapter 5

People detection based on appearance and motion information

5.1 Introduction[1]

Nowadays, many different systems exist which try to solve the people detection problem. The state of the art in people detection and tracking includes several successful solutions working in specific and constrained scenarios. However, the people detection complexity is higher in real world scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability. Over the last few years, there have been multiple approaches in more realistic environments with multiple people and occlusions [Andriluka et al., 2008; Leibe et al., 2005] and even onboard scenarios [Wojek et al., 2009]. Most of them get acceptable results using only the appearance information or adding tracking information. As already commented in the previous chapter 4, due to the difficulty to obtain a reliable segmentation in more complex and realistic scenarios, people detection approaches based on segmentation does not work properly in this kind of scenarios. The main objective of this chapter is to present a new people detection approach based on motion and the combination of appearance and motion information in order to achieve a more reliable performance and work in more complex and realistic scenarios.

In this chapter, we will firstly make a brief introduction to the related literature in section 5.2. Then, the proposed appearance and motion people detection approach is described in section 5.3. After that, section 5.4 describes the experimental results. Finally, section 5.5 summarizes the chapter with some conclusions.

[1] This chapter is based on the publication *"A. García-Martín, A. Hauptmann, J. M. Martínez. People detection based on appearance and motion models. In Proc. of the IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 256–260, 2011"*

5.2 Related work

In the following, we give an overview of current people detection approaches focusing on the kind of information they employ: appearance and/or motion. As already described in chapter 2, there is a more comprehensive study of the use of appearance information in the state of the art, mainly due to the fact that appearance provides a much more discriminant information about people detection. There are some approaches that include motion information to add robustness to the detection and there are very few cases where the only information used is motion.

Those approaches based on appearance information can be classified in two major groups. On the one hand, the methods based on simplified person or holistic models (only a region or shape): [Xu and Fujimura, 2003] uses an ellipse model and a silhouette fitting algorithm, [Zhou and Hoang, 2005] performs the classification by similarity with silhouettes stored in a codebook, [Dalal and Triggs, 2005] uses a person model based on the Histogram of Oriented Gradients (HOG) descriptors and a Support Vector Machine (SVM) classifier and [Leibe et al., 2008] makes use of shape representation with the generative Implicit Shape Model (ISM) framework. On the other hand, there are methods based on combination of multiple parts or part-based models: [Andriluka et al., 2009] trains multiple detectors for anatomically defined body parts which are then combined using pictorial structures, [Haritaoglu et al., 1998] performs an analysis of concavity and convexity of the silhouette to identify different body parts, [Wu and Nevatia, 2005] tries to identify the characteristic edges of a human body and to generate four edge models (body, head, torso and legs) independently trained using a nested Adaboost cascade structure and in chapter 4 a real time adaptation of the work described in [Wu and Nevatia, 2005] has been presented.

Although it is known that human motion is an important cue for people detection, there are not many approaches that make use of this information. Some authors combine appearance and motion expanding their own previous works to more than one frame [Viola et al., 2003; Dalal and Triggs, 2006], improving the results significantly, but without generating a motion model as an independent entity. Some approaches use only the motion information: [Cutler and Davis, 2000] applies time-frequency analysis to detect and characterize the human periodic motion and [Sidenbladh, 2004] detects patterns of human motion using optical flow and an SVM classifier.

We propose a people detector working in more complex and realistic scenarios. It extracts the objects candidates to be a person with exhaustive search, allowing more robustness to multiples scales and rotations. However, the exhaustive search and the feature extraction require high computational cost, so it does not works in real time. The main contribution presented in this chapter is a new motion model inspired by the well-established ISM people detection approach [Leibe et al., 2008] and the MoSIFT descriptor [Chen and Hauptmann, 2009], which has been successfully employed in activity recognition. Combining both ideas, a new people detection approach based on their motion is introduced: Implicit Motion Model (IMM). Furthermore,

to evaluate this new detector, a full system that combines appearance, motion and tracking information has been designed and developed.

5.3 People detection based on appearance and motion models

5.3.1 System overview

As already discussed in chapter 2, the basic architecture of any people detector includes: an image or video input, the object detection task, the designed and trained (if training is required) person model, the verification/classification task and the final detection decision (see Figure 2.1). In order to evaluate our proposal, a complete framework has been designed to predict or update the visual people detection (see Figure 5.1). It is able to perform two independent visual people detections, the first one using the shape or appearance of humans as discriminative feature and the second one using their motion. Using the people detection as first step, the framework is able to update the person detection (appearance, motion or their fusion) iteratively over time using a color based tracker.

Input

The system is based on color frames extracted from static and mono camera video sequences.

Object detection

The object detection technique is feature-based exhaustive search. It is a commonly used technique for localizing the objects candidates to be a person in the scene. In our system, feature-based exhaustive search is based on [Leibe et al., 2008]. It builds up the detection confidence density over the full image explicitly in a bottom-up fashion through probabilistic votes cast by local features matching.

Person model

The chosen person model corresponds to the combination of appearance and motion information. A person model based on appearance and another person model based on motion are defined independently. The details of this module in our system are described with more detail in section 5.3.2.

Verification/Classification

This process compares the previously trained appearance and motion person models with the candidates to be person from an image or sequence and combines the results of both models. The details of this module in our system are described with more detail in section 5.3.2.3.

Object tracking

In our system, tracking is based on [Nummiaro et al., 2003], adding to the particle filter algorithm an adaptive appearance model based on color distributions. The object model is represented by a weighted histogram which takes into account both the color and the shape of the target. It also includes a straightforward kinematic system model to propagate the particle

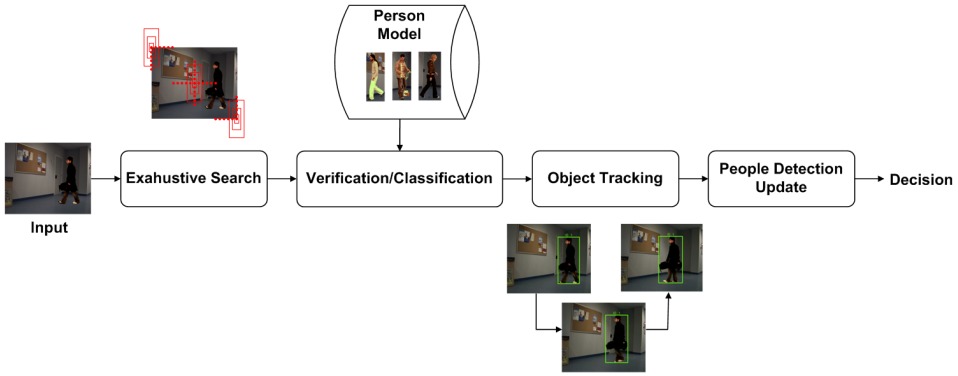


Fig. 5.1. Overall system architecture.

filter sample set. The observation probability of the particle filter mean state will be used as tracker confidence level C_t^{track} in the people detection update.

People detection update

A tracking process is initialized for each detected person. The following detections will update existing trackers or will create new tracking processes. The conditional probability of people detection, given the tracking information in each frame $P_t^{det|track}$, will be predicted or updated over time based on current people detection probability P_t^{det} and the tracker confidence level C_t^{track}:

$$P_t^{det|track} = \begin{cases} P_t^{det}, & P_t^{det} > 0 \\ P_{t-1}^{det|track} - (1 - C_t^{track}), & P_t^{det} = 0 \end{cases} \tag{5.1}$$

Decision

For performance evaluation methodology purposes (see section 3.3), the output detection confidence is evaluated in terms of Precision and Recall at the operating point, i.e., at the predefined optimal decision threshold for each algorithm (score threshold decision) as well as the AUC-PR to summarize the overall performance.

5.3.2 People detection approach

The proposed people detector is able to perform two independent visual people detections, the first one using the shape or appearance of humans as discriminative feature and the second one using their motion. The final detector is the combination of both sources.

5.3.2.1 Appearance person model

The appearance people detector is based on the Implicit Shape Model (ISM) [Leibe et al., 2008]. ISM is a generative model for object detection and has been applied to a variety of object

categories including cars, motorbikes, animals and pedestrians. The ISM consists of a codebook C_{ISM} of local appearances, that are prototypical for the object category and a spatial probability distribution $P^{C_{ISM}}$ which specifies where each codebook entry may be found on the object. The K elements of C_{ISM} are local descriptors $d_1^{C_{ISM}}, ..., d_K^{C_{ISM}}$ extracted around scale-invariant interest points (x_k, y_k, s_k). The codebook C_{ISM} is generated using an agglomerative clustering with average linkage and only the cluster centers are stored. The spatial probability distribution $P^{C_{ISM}}$ is learned during a second training phase where all the local descriptors are matched in multiple clusters with different weights.

5.3.2.2 Motion person model

The pattern of human motion is well known to be really discriminative from other types of motions [Cutler and Davis, 2000; Dalal and Triggs, 2006; Viola et al., 2003]. We introduce a new human motion representation that is mainly based on the use of the ISM framework [Leibe et al., 2008] and the motion information in the MoSIFT descriptor [Chen and Hauptmann, 2009].

MoSIFT MoSIFT [Chen and Hauptmann, 2009] is a variation of the well-known SIFT point detector and descriptor [Lowe, 2004]. MoSIFT detects interest points and encodes not only their local appearance, but also their explicit local motion. It consists of three main steps: firstly, the SIFT algorithm is applied to find scale-invariant interest points in the spatial domain, then optical flow is extracted around the distinctive points with (temporal) motion constrains at corresponding scales and, finally, the feature descriptor is generated. Figure 5.2 shows results of SIFT and MoSIFT over the same frame.

In order to generate the feature descriptor, MoSIFT adapts the idea of grid aggregation in SIFT to describe motion, but instead of using appearance gradients, it uses the optical flow. The other main difference to appearance description is in the rotation invariance. Rotation invariance is important to appearance since it provides a standard to measure the similarity of two interest points, but the direction of movement is actually an important (non-invariant) vector to discriminate different movements. The two aggregated histograms (appearance and optical flow) are combined into the MoSIFT descriptor, which has therefore 256 (128+128) dimensions.

Implicit Motion Model The main idea consists of identifying and learning characteristic motions of humans in typical surveillance systems and generating a motion model. We propose to use the motion information in the MoSIFT descriptor to characterize the movements and build a motion model following the ISM framework.

For symmetry with the ISM model, the IMM consists of a codebook C_{IMM} of local motions, that are prototypical for the object category, and a spatial probability distribution $P^{C_{IMM}}$ which specifies where each codebook entry may be found on the object. The K elements of C_{IMM} are

Fig. 5.2. SIFT (left) and MoSIFT (right) interest points. Yellow circles indicate interest points and their scales, red arrows indicate the dominant motion orientation.

the motion part of MoSIFT [Chen and Hauptmann, 2009] descriptors $d_1^{C_{IMM}}, ..., d_K^{C_{IMM}}$ extracted around scale-invariant and spatio-temporal interest points (x_k^t, y_k^t, s_k^t). The codebook C_{IMM} is generated using the Reciprocal Nearest Neighbors (RNN) clustering algorithm [Leibe et al., 2008] and the spatial probability distribution $P^{C_{IMM}}$ is learned using annotated training sequences or pairs of images; our training dataset includes several sequences, but other datasets only include pairs of images which are enough for training the motion model.

5.3.2.3 Verification/Classification

Given a new test pair of images, the SIFT interest point detector is applied, then SIFT and MoSIFT features are extracted around the selected locations. These features are matched to the corresponding learned codebook C_{ISM}/C_{IMM} in multiple clusters with different weights. Each matching casts votes for theoretical positions of the person center according to the corresponding learned spatial distribution $P^{C_{ISM}}/P^{C_{IMM}}$. Then, the hypotheses are defined as local maxima in the voting space (x, y, s). Assuming symmetry with respect to our hypothetical centers, a bounding box is obtained for each hypothesis. Finally, multiple hypotheses with more than 50% cover and overlap, as defined in [Leibe et al., 2005], are simplified to the highest score one. Figure 5.3 shows two IMM detection examples of the same sequence.

Frame by frame, each detector generates a list of blobs, B_t^{ISM} or B_t^{IMM}, with the associated people detection probability, P_t^{ISM} or P_t^{IMM}. The appearance and motion detectors have been combined at blob level (position and dimension). It means that both detectors have been run independently and the results (blobs) have been considered as multiple detection hypotheses. Finally, multiple hypotheses with more than 50% cover and overlap and less than 0.5 relative distance between centers (Multiple Hypotheses Simplification Criteria, $MHSC$) are simplified

Fig. 5.3. IMM detection process examples. Voting space (black lines), center hypotheses (green points), hypotheses (red rectangles) and final hypothesis (green rectangles).

to the average hypothesis: blob B_t^{det} and detection probability P_t^{det} (see algorithm 5.1).

5.4 Experimental results

This section describes the experimental dataset (training and test dataset), the results obtained in each stage of our system and the computational cost.

5.4.1 Experimental setup

Focused on the idea of evaluating the performance of the proposed approach in more complex and realistic scenarios, we use the evaluation dataset B of the PDds dataset (see section 3.3.1).

In order to train the people motion model, the evaluation dataset has been divided in training and test. To be homogeneous, the detector and the tracking approach have been evaluated on the same video sequences: the test dataset composed of 36 sequences[2]. All persons manually annotated at the scene have been taken into account in the evaluation.

The training dataset is composed of the other 25 sequences. Each sequence includes multiple annotated people, but the IMM has been trained using only the *MMI person*: the person with Maximum Motion Information (MMI) per video. The *MMI persons* have a trajectory completely non-occluded since entering the scene until they come out of it. The *MMI persons* have been manually selected in each video. The 25 training sequences have been selected in order to contain as many different cases (directions, scales, etc) as possible.

[2]Test sequences (referring to PDds numbering): 2-5, 7-8, 12, 14, 18, 32, 34, 36-38 and 40-61.

Algorithm 5.1 Appearance-Motion blobs combination.

1. Number of final hypotheses $k = 0$.
2. For $i = 1$ to $card\left(B_t^{ISM}\right)$.
   (a) For $j = 1$ to $card\left(B_t^{IMM}\right)$.
      i. If $MHSC\left(B_t^{ISM}(i), B_t^{IMM}(j)\right) = true$.
      $k = k + 1$.
      $B_t^{det}(k) = \frac{B_t^{ISM}(i) + B_t^{IMM}(j)}{2}$.
      $P_t^{det}(k) = \frac{P_t^{ISM}(i) + P_t^{IMM}(j)}{2}$.
3. For $r = 1$ to remaining blobs (B_t^{ISM} or B_t^{IMM}).
   $k = k + 1$.
   $B_t^{det}(k) = B_t^{ISM}(r)$ *or* $B_t^{IMM}(r)$.
   $P_t^{det}(k) = P_t^{ISM}(r)$ *or* $P_t^{IMM}(r)$.

5.4.2 People detection results

In order to evaluate the different people detectors and the integrated system, firstly we have evaluated each separate detector and their fusion over the 36 test sequences. The appearance and motion detectors have been combined at blob level: both detectors have been run independently and the results (blobs) have been added, or have been averaged in those cases of overlapping blobs (see section 5.3.2.3). The ISM results have been obtained using the available code and binaries[3] and the IMM has been implemented using the LIBPMK library [Lee, 2008] . The ISM has been already evaluated in the dataset B in terms of AUC-PR (see section 4.4.2.2). In this chapter, trying to observe more significant differences between the approaches, it has been decided to evaluate the operational performance in a real or final system, i.e., at the operating point (see section 3.3.2).

We can see in Table 5.1 the average results for the test data (Precision, Recall and F1Score). We can see how both algorithms with high Precision values (∼94%) differ in Recall values (12∼16%). It is logical that the motion-based detector obtains lower Recall values because only moving people can be detected. However, in environments as complex as these ones, the use of motion information obtains results close to the use of appearance information. The combination of both detectors obtains better Recall results (21.7%), slightly reducing Precision values (93.9%).

Secondly, we have evaluated the whole system over the same 36 test sequences. Using algorithms with high Precision values (∼94%), our prediction or update based on tracking confidence is able to maintain high Precision values (91.8∼93.7%), but improving considerably the Recall

[3] `http://www.vision.ee.ethz.ch/~bleibe/index.html`

Approach	Precision	%Δ	Recall	%Δ	F1Score	%Δ
ISM	94.7	0.0	16.5	0.0	27.2	0.0
IMM	95.1	+0.4	12.1	-26.7	21.2	-22.1
ISM+IMM	93.9	-0.8	21.7	+31.5	34.6	+27.2

Table 5.1: Detection results. Percentage increase (%Δ) calculated with respect to ISM.

Approach	Precision	%Δ	Recall	%Δ	F1Score	%Δ
ISM+Tracking	93.7	-1.1	22.8	+38.2	37.4	+37.5
IMM+Tracking	93.1	-2.1	18.6	+53.7	32.1	+51.4
ISM+IMM+Tracking	91.8	-2.2	28.4	+30.9	44.6	+28.9

Table 5.2: System results. Percentage increase (%Δ) calculated with respect to each approach without tracking.

(18.6∼28.4%). We try to correct the people detection "unstable" behavior over time with the tracking information. The people detection probability prediction and update allows us to stabilize the detection over time and to eliminate false positive detections quickly. We can see in Table 5.2 the average results of three different system configurations, the ISM detector, the IMM detector and their fusion, all of them adding the tracking information.

Every video surveillance system and people detector must maintain a compromise between Precision and Recall. Thinking about the people detection as a preliminary step in the event detection task (e.g., TRECVID Surveillance event detection), it is more valuable to get better Recall results at the expense of getting slightly reduced Precision results. At higher semantic levels (activity recognition or detection), the people detection false positives can be easily dismissed, but on the other hand the undetected people cannot be recovered.

5.4.3 Computational cost

According to the computational cost, the IMM detector is based on the ISM Framework, but with the MoSIFT features instead of the SIFT features. The use of MoSIFT features increases the final computational cost due to the optical flow computation, but in the case of MoSIFT features, the number of features to be processed after the feature extraction is highly reduced. Unless the computational cost of the optical flow, both detectors (ISM and IMM) have comparable computational costs. According to the original ISM approach [Leibe et al., 2008], typical runtimes of the pedestrian detector range between 4-7 seconds with 320x240 images. We expect that the IMM performance can still be considerably improved by a more efficient implementation of the optical flow (being already an available real time implementation on OpenCV[4]). Even so,

[4] `http://sourceforge.net/projects/opencv/`

the computational cost is still far from the real time operation.

Finally, running both people detectors in parallel and being almost insignificant the computational cost of the combination in comparison with the computational cost of the detectors, the final combination approach computational cost will be established by the detection approach with the higher computational cost, i.e., the IMM.

5.5 Summary and conclusions

In this chapter, a new people detection motion model IMM is proposed. Using the ISM Framework and the MoSIFT interest points detector and descriptor, we present a new people detection algorithm taking into consideration the motion of people. It is clear that human motion provides useful information for people detection and independent from appearance information, so we also present an integrated system which combines an appearance people model, our new motion model and a tracking algorithm. Experiments have been conducted on challenging and realistic sequences extracted from the TRECVID dataset that are part of our evaluation dataset PDds with the maximum complexity category C5 (see chapter 3.3.1). The results show that our motion-based detector produces results comparable to the ISM state of the art approach in complex and realistic scenarios. The evaluation of the whole system shows how the combination of different information sources improves the final detection, obtaining a significant improvement in Recall and a slightly Precision reduction.

In the following chapter, we take advantage of this appearance and motion combination over time. So that the improvement introduced by tracking in the detection at one frame becomes a potential detection and tracking improvement in the following frames and vice versa.

Chapter 6

Collaborative people detection and tracking

6.1 Introduction[1]

As already mentioned, people detection is one of the most challenging problems in computer vision. People detection approaches from the state of the art obtain satisfactory results in low and medium complexity scenarios, but these results are considerably reduced in more complex and realistic scenarios (see chapter 4). In order to achieve a more reliable performance in complex scenarios, we have proposed a new people detection approach based on motion and the combination of appearance and motion information (see chapter 5). In this chapter, we propose the integration of the appearance and motion information in a detection and tracking system that takes advantage of the tracking information, improving the detection results over time.

Object video tracking is the process of locating a moving object (or multiple objects) over time using information extracted from a video sequence. Tracking is one of the main tasks in video analysis, being essential in a multitude of tasks such as video surveillance, traffic monitoring, vehicle navigation, human-machine interaction, etc. Traditionally, the performance of object tracking systems is based only on low-level visual or motion characteristics of the target objects: points, color, shape, speed, etc. We propose the use of people detection not only as tracker initialization, but also as an additional higher-level tracker input feature in order to dynamically update the tracker with each new object detection, improving the overall tracking results. We demonstrate this proposal using persons as our objects of interest, but this may be adapted to other kind of objects.

[1] This chapter is based on the publications *"A. García-Martín, J. M. Martínez. On collaborative people detection and tracking in complex scenarios. Image and Vision Computing, 30 (4-5):345-354, May 2012"* and *"A. García-Martín, J. M. Martínez. Enhanced people detection combining appearance and motion information. Electronic Letters, 49 (4): 256-258, January 2013"*

The main contribution presented in this chapter is a detection/tracking collaborative scheme that integrates appearance, motion and tracking information. Each task follows a parallel process and provides useful information to the other process frame by frame. The collaborative system consists of successive stages of information exchange, so the improvement introduced by one process becomes a potential self-improvement in the following stages.

In this chapter, we will firstly make a brief introduction to the related literature in section 6.2. Then, the proposed detection/tracking collaborative system which combines different information sources is described in section 6.3. After that, section 6.4 describes the experimental results. Finally, section 6.5 summarizes the chapter with some conclusions.

6.2 Related work

As discussed previously, this chapter is focused on two of the most common tasks performed by video surveillance systems: people detection and tracking. The people detection state of the art has been widely described in chapter 2 and a more thoroughly description of the difference between detectors based on appearance or motion information was done in the previous chapter 5. For this reason, the following sections include a brief state of the art of tracking and its combination with people detection.

6.2.1 Tracking

Multiple techniques have been developed for object tracking. Every tracker must define its target object representation, the features used to define this object and how to parametrize the evolution of the object over time. In [Yilmaz et al., 2006], according to the object representation, the authors have classified the object tracking methods into three categories: point tracking, kernel tracking and silhouette tracking. The object representations traditionally used are points, shapes, silhouettes, etc; and the chosen features are color, edges, optical flow, etc or combinations of them.

Many tracking approaches use the color information for its great discriminatory power: for example, [Comaniciu et al., 2003] uses a color histogram with an isotropic kernel as object model and [Nummiaro et al., 2003] uses a color histogram with an adaptive particle filter. [Zhou et al., 2009] uses the edge information about the SIFT features and the Mean Shift algorithm. There are some approaches that use motion information, for example optical flow [Cremers and Schnörr, 2003; Denman et al., 2007], or even textures [Jepson et al., 2003]. And there are many approaches that try to combine some of the previous mentioned features in order to improve the final tracking results (e.g. [Pérez et al., 2004; Wang and Yagi, 2008]).

6.2.2 Detection and tracking combination

Most approaches from the state of the art that combine detection and tracking are designed mainly with the aim of improving tracking results (tracking-by-detection) [Leibe et al., 2007; Andriluka et al., 2008; Ess et al., 2009; Stalder et al., 2010; Yu et al., 2011; Wu and Nevatia, 2007; Ren, 2008; Giebel et al., 2004; Okuma et al., 2004; Avidan, 2007; Li et al., 2008; Breitenstein et al., 2010] and the improvements introduced in the detection task are a byproduct of the tracking task (detection-by-tracking), i.e., the tracking results are assumed by default as improved detection results. Table 6.1 summarizes the different approaches from the state of the art, the methods used and the evaluation for detection, tracking, detection-by-tracking and tracking-by-detection. Some approaches extract overcomplete sets of trajectories or tracklets using detection information and, then, they make a global optimization in order to prune and select the final trajectories [Leibe et al., 2007; Andriluka et al., 2008; Ess et al., 2009; Stalder et al., 2010; Yu et al., 2011]. Other approaches perform directly data-matching and linking between detections in order to solve the tracking problem [Wu and Nevatia, 2007; Ren, 2008; Avidan, 2007]. Finally, some approaches (including our approach) combine detections and particle filtering results [Giebel et al., 2004; Okuma et al., 2004; Li et al., 2008; Breitenstein et al., 2010] making use of the detection information in order to guide or weight each particle filter iteration.

Within the state of the art, the most similar approaches to our proposal are those that, in addition to the tracking enhancement, try to improve or update explicitly the detection using the tracking history (detection-by-tracking) [Avidan, 2007; Li et al., 2008; Breitenstein et al., 2010]. [Avidan, 2007] presents a tracking approach based on an online people detector and the Mean Shift algorithm, [Li et al., 2008] combines a particle filter tracker with three different observers (online face detectors) where each observer is learned or updated with a different subset of previous faces samples and [Breitenstein et al., 2010] proposes a particle filter tracker with an observation model based on three detection components: a high confidence people detection, a continuous confidence detection and an online detection. However, none of these approaches are able to perform both tasks independently and none of them makes use of the tracking confidence level: they only make use of the tracker trajectory. In addition, they are focused on tracking; the main objective of improving the detection is in order to improve the tracking, so they do not even evaluate the detection process.

The benefits of using the detection information to improve the tracking have been already reported in the state of the art, but we also introduce the opposite flow of information to improve the detection, so that we define a collaborative scheme system that integrates the people detection and tracking information into a single system and improves both tasks simultaneously (detection-by-tracking and tracking-by-detection). State of the art people detectors usually get high Precision results, but they are not stable over time, i.e., they have an intermittent performance. A person can be detected in an instant of time and not be detected in the next instant

Approach	Detection				Tracking				Evaluation dataset[1]		
	Baseline	Eval.	by-Tracking	Eval.	Baseline	Eval.	by-Detection	Eval.	Detection	Tracking	Complexity
[Leibe et al., 2007]	ISM	Yes	By default	Yes	Tracklets	Yes	Yes	Yes	–1 seq–		Complex
[Andriluka et al., 2008]	Pictorial structures	Yes	By default	Yes	Tracklets	No	Yes	No	2 seq	-	Complex
[Ess et al., 2009]	ISM	Yes	By default	Yes	Tracklets	No	Yes	Yes[2]	3 seq	4 seq	Complex
[Stalder et al., 2010]	HOG	Yes	By default	Yes	Tracklets	No	Yes	Yes[2]	2 seq	2 seq	Medium
[Yu et al., 2011]	Adaboost	Yes	By default	Yes	Tracklets	No	Yes	Yes	2 seq	6 seq	Complex
[Wu and Nevatia, 2007]	Edge body parts	Yes	By default	Yes	Mean-Shift	No	Yes	Yes	1 seq	3 seq	Medium
[Ren, 2008]	Adaboost	Yes	By default	Yes	Kalman Filter	No	Yes	No	3 seq	-	Complex
[Giebel et al., 2004]	Silhouette matching	No	By default	No	Particle filter	No	Yes	Yes[2]	-	3 seq	Simple
[Okuma et al., 2004]	Adaboost	Yes[2]	By default	Yes[2]	Particle filter	No	Yes	Yes[2]	–1 seq–		Medium
[Avidan, 2007]	Online Adaboost	No	Yes[3]	No	Mean-Shift	Yes[2]	Yes	Yes[2]	-	9 seq	Simple
[Li et al., 2008]	Online Adaboost	No	Yes[3]	No	Particle filter	No	Yes	Yes	-	6	Medium
[Breitenstein et al., 2010]	ISM/HOG	No	Yes[3]	No	Particle filter	No	Yes	Yes	-	10 seq	Complex
Ours	ISM/IMM/HOG/ TUD/DTDP	Yes	Yes	Yes	Particle filter	Yes	Yes	Yes	–36 seq–		Complex

Table 6.1: Detection and tracking combination approaches from the state of the art. The table summarizes the methods used and the evaluation for detection, tracking, detection-by-tracking and tracking-by-detection. [1]Number and complexity of sequences used for the evaluation. [2]Only qualitative evaluation. [3] Includes detector online update, but without the explicit use of the tracking confidence level (only trajectory).

of time even with a minimal variation of appearance (consecutive frames). With the tracking information, we are able to extrapolate in time these intermittent detections that a priori are clearly undetected people. Some approaches do this extrapolation by default, that is, without taking into account additional information sources (only trajectory). In our case, the additional information from tracking (confidence level) provides an external evidence supporting the removal of the detector errors (or supporting its decision). In turn, the tracking information allows us to dismiss quickly this extrapolation in time according to a fast tracker confidence decrease (person going out of the scene or occlusions).

Our system is designed with the aim of evaluating each task independently and their combination. Although, the system evaluates different people detectors with a particle filter tracker, the detection and tracking modules can be replaced by others from the state of the art without great difficulty. The use of different modules will vary the overall performance of the system, but the combination of both sources of information will always be useful for improving the system (except in the ideal case of perfect detection and perfect tracking).

6.3 People Detection/Tracking collaborative system

This section describes the integration of our people detector based on the combination of appearance and motion information (see previous chapter, section 5.3.2) in a complete people detection and tracking system.

6.3.1 System overview

A complete people detection and tracking system has been designed (see Figure 6.1). The system framework has not only been designed to perform each task independently (after tracking initialization), but also to support the transfer and use of mutual information between them. Each task (detection and tracking) follows a parallel process and provides useful information to the other process frame by frame. In order to facilitate the evaluation of any detection and tracking algorithm from the state of the art in our framework: firstly, the format of the mutual exchanged information has been designed as generic as possible and, secondly, the exchange mechanism has been designed to be easily compatible and without requiring any prior training or modeling (as they are very dependent on the chosen approaches and scenario).

The detection process, whether based on the appearance, movement, tracking or any combination of them, always consists of a list of detections in each frame. Each detection is represented by its position, dimensions (bounding box) and people detection probability. This information is used by the tracker processes to update their target models (or to initialize them).

The tracking process consists of a set of tracker instances. Each tracker instance needs to be initialized, so if there is a new detection that is not associated with any tracker instance, a

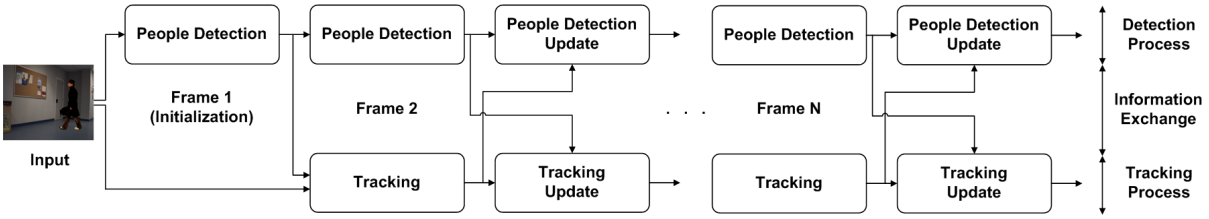


Fig. 6.1. Overall system architecture.

new tracker instance is created, initialized and associated with this detection. The association maximizes the cover and overlap between blobs and follows the evaluation criteria defined in [Leibe et al., 2005] (relative distance between centers, cover and overlap). Finally, each tracker instance result is represented by a trajectory of the associated object or person in the scene and each trajectory is described at each instant of time by a blob and the associated tracker confidence level. This information is used by the people detector to update the person detection probability over time.

Frame by frame, the information generated by each task is used as update or prediction information for the other. In the case of no people detections or if the tracker is unable to track its target, each process is able to continue its operation independently of the other. Therefore, the system supports three different system configurations: detection, tracking and collaborative detection-tracking. In the following sections, we describe the performance of the people detection and tracking modules of our system and the associated update processes.

6.3.2 People detection

The proposed people detector is already described in previous chapter (see section 5.3.2). It is able to perform two independent visual people detections, the first one using the shape or appearance of humans as discriminative feature and the second one using their motion. The final detector is the combination of both sources.

6.3.3 Tracking

The basic tracker is based on [Nummiaro et al., 2003] where an adaptive appearance model based on color distributions is added to the particle filter algorithm, also including a target model update process in order to support object variations (pose, illumination, camera point of view, etc). The particle filter is based on a weighted sample set $S = \{(s(n), \pi(n)) \mid n = 1, ..., N\}$, where each sample s represents one hypothetical state of the object (position, dimension and velocity) with a corresponding discrete sampling probability π, where $\sum_{n=1}^{N} \pi(n) = 1$. Then the

mean state of an object is estimated at each time by

$$E[S] = \sum_{n=1}^{N} \pi(n)\, s(n) \tag{6.1}$$

The object model is represented by the state s and the associated weighted histogram (m bins) which takes into account the object color distribution. The similarity measure used between the target q and any $h(n)$ color distribution of the N hypotheses is the Bhattacharya coefficient ρ and the corresponding Bhattacharya distance:

$$d(n) = \sqrt{1 - \rho[h(n), q]},\ n = 1, ..., N \tag{6.2}$$

The tracker employs the Bhattacharya distance to update the a priori distribution calculated by the particle filter, so that small Bhattacharya distances correspond to large weights that are specified by Gaussian distribution with variance $\sigma = 0.1$:

$$\pi(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{d(n)^2}{2\sigma^2}} \tag{6.3}$$

To allow target model variations, either by extrinsic (e.g., lighting changes, point of view variations, occlusions) or intrinsic object (e.g., pose, clothes variations) factors, it performs a target color distributions update with the color histogram of the particle filter mean state $h_t^{E[S]}$ at each instant according to an update factor α:

$$q_t = (1 - \alpha)\, q_{t-1} + \alpha h_t^{E[S]} \tag{6.4}$$

6.3.4 Update of people detection and tracking modules

Through the information exchange between the processes, the people detection and the people tracking modules are capable of using the information from the other one in order to self-correct or self-update. This section describes the people detection and tracking update modules which allow the third system configuration: the collaborative detection-tracking.

6.3.4.1 People detection update

Using the people detection as first step, the global detection process is able to update the person detection (appearance, motion or their fusion) iteratively over time using the tracking information (see Figure 6.2). A new tracker instance is initialized for each new detected person, not already associated with any other tracker instance. The following detections will be associated with existing trackers or will create new tracker instances. These trackers allow us to update the final people detection over time. In those trackers with new associated detections, the

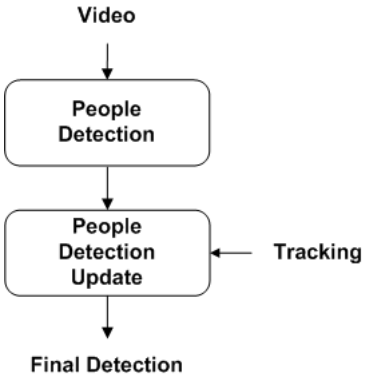


Fig. 6.2. People detection update. The people detection is updated using the tracking information.

detection blob is used directly to update the corresponding tracker instance with a high update factor β (see section 6.3.4.2), so the corresponding tracker instance confidence level will be always high and, therefore, not discriminatory or reliable. On the contrary, in those trackers without new associated detections, the corresponding tracker instance is updated with a detection estimation that follows two steps: in the first step, the people detection is predicted as an averaged detection over time and, then, in the second step, this prediction is updated according to the tracker confidence level. Finally the conditional probability of people detection given the tracking information in each frame, $P_t^{det|track}$, is predicted or updated over time based on the current people detection probability P_t^{det} (obtained by the people detection module, see section 5.3.2.3) or the cumulative conditional probability over time and the tracker confidence level C_t^{track}(observation probability of the mean state $\pi^{E[S]}$[Nummiaro et al., 2003]):

$$P_t^{det|track} = \begin{cases} P_t^{det}, & P_t^{det} > 0 \\ \frac{1}{t-1}\sum_{i=1}^{t-1} P_i^{det|track} - (1 - C_t^{track}), & P_t^{det} = 0 \end{cases} \quad (6.5)$$

6.3.4.2 Tracking update

The main idea of our proposed tracking process consists of using the adaptive tracking algorithm proposed in [Nummiaro et al., 2003] and adding the people detection described previously as an additional information source (see Figure 6.3).

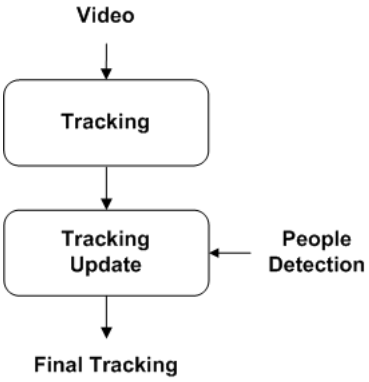


Fig. 6.3. Tracking update. The tracking is updated using the people detection information.

As already mentioned, a new tracker instance is initialized for each new detected person. In absence of following people detections associated with a tracker, each tracker will continue with its normal operation, but following associated detections will update the respective tracker exploiting the capability for updating the target model (see Equation 6.4). We propose to add a new update stage with the people detection information: with each new associated people detection, the detected blob is used to calculate a new color distribution that is considered as a new hypothesis h_t^{det} that updates the target model by a factor β:

$$q_t^{det} = (1-\beta)\, q_t + \beta h_t^{det} \tag{6.6}$$

Finally, the existence of people detections (current conditional probability of people detection $P_t^{det|track}$) controls the entire upgrade process of the target model (q_t):

$$q_t = \begin{cases} (1-\alpha)\, q_{t-1} + \alpha h_t^{E[S]}, & P_t^{det|track} = 0 \\ (1-\beta)\left((1-\alpha)\, q_{t-1} + \alpha h_t^{E[S]}\right) + \beta h_t^{det}, & P_t^{det|track} > 0 \end{cases} \tag{6.7}$$

6.4 Experimental results

This section describes the experimental setup used in the evaluation of the proposed people detection and tracking approaches, the results of the evaluation of each system configuration (detection, tracking and collaborative detection-tracking) and the computational cost.

6.4.1 Experimental setup

As in the previous chapter, focused on the idea of evaluating the performance of the proposed approach in more complex and realistic scenarios, we use the evaluation dataset B (see section 3.3.1). It has also used the same division of the training and test sequences (see section 5.4.1).

As already described, the training dataset has been used only to train the people motion model of the IMM approach.

6.4.2 People detection results

This section describes the results of the people detection experiments over our test dataset. It includes a comparison of four different appearance approaches from the state of the art (ISM [Leibe et al., 2008], HOG [Dalal and Triggs, 2005], TUD [Andriluka et al., 2009] and DTDP [Felzenszwalb et al., 2010]), the already described motion approach IMM (see previous chapter 5) and all the appearance-motion combinations (ISM+IMM, HOG+IMM, TUD+IMM and DTDP+IMM). The combination of ISM and IMM has been described in detail in previous chapter 5. In this chapter, the other three appearance-motion combinations have been introduced with similar performance patterns. The different approaches from the state of the art (ISM, HOG, TUD and DTDP) have been already evaluated in the dataset B in terms of AUC-PR (see section 4.4.2.2). As in the previous chapter, trying to observe more significant differences between the different approaches, it has been decided to evaluate the operational performance in a real or final system, i.e., at the operating point (see section 3.3.2).

In order to evaluate the different people detectors, firstly we have evaluated (Precision, Recall and F1Score) each separate detector and their fusion over the 36 test sequences.

Table 6.2 shows the average results for the test dataset. We can see how in general all algorithms have high Precision values (93∼96%) and low Recall values (10∼26%). It is due mainly to two reasons: in first place, the content set contains challenging scenarios, crowds and a wide range of scales. It is easier to detect people with higher scales and without occlusions (better visual and motion information); secondly, the people detectors have high Precision values, but there are not stable over time, that is, a person could be detected in one frame and could be not detected in the next one, even when the difference between consecutive frames is minimal.

The fusion of two independent information sources (appearance and motion) provides a significant improvement in terms of Recall (31∼58%) and F1Score (27∼49%), without a significant Precision variation. Finally, the detectors ISM+IMM, HOG+IMM and DTDP+IMM get similar results, whilst the detector TUD+IMM is clearly worse.

6.4.3 Tracking results

This section describes the results of the tracking experiments over the test dataset. Each sequence includes multiple annotated people: all persons detected in the scene have been tracked. As already explained previously, a tracker is initialized for each detected person and the final video result is the average of all people tracked.

To compare the performance of different approaches, we have calculated the average tracking Precision, Recall and F1Score over time in terms of blob overlapping between the hypothesis

Approach	Precision	%Δ	Recall	%Δ	F1Score	%Δ
ISM	94.7	-	16.5	-	27.2	-
HOG	93.4	-	15.4	-	25.3	-
TUD	95.1	-	10.7	-	19.1	-
DTDP	95.1		19.5		30.7	
IMM	95.1	-	12.1	-	21.2	-
ISM+IMM	93.9	-0.8	21.7	+31.5	34.6	+27.2
HOG+IMM	96.4	+3.2	21.5	+39.6	34.5	+36.4
TUD+IMM	94.4	-0.7	17.0	+58.9	28.5	+49.2
DTDP+IMM	96.4	+1.4	26.3	+34.9	39.2	+27.7

Table 6.2: Detection results. Percentage increase (%Δ) calculated with respect to single appearance versions.

and the ground-truth blobs. The Precision is the number of pixels that are correctly detected as belonging to the target vs. the total number of tracked pixels, whilst the Recall is the number of pixels that are correctly detected as belonging to the target vs. the total number of ground-truth pixels. The DScore has been computed relative to the object size according to [Leibe et al., 2005] and the LostRate (percentage of time that the target is lost) has been also calculated. All these measures have been averaged in all the test videos. The experiments include the evaluation of the original tracking algorithm [Nummiaro et al., 2003] with different people detection initializations, namely: ISM, HOG, TUD, DTDP, IMM, ISM+IMM, HOG+IMM, TUD+IMM and DTDP+IMM.

The approach has been evaluated with different variations of the update parameter ($0 \leq \alpha \leq 1$), obtaining a common pattern (see Figure 6.4). To discuss the results in Table 6.3, the update parameter of the proposed method has been fixed to $\alpha=0.2$. We can see how all trackers with different initializations have similar results in terms of F1Score (39∼50%). The HOG generates bigger blobs than the other detectors, so it gets the best Recall results (78.0%), but the worse Precision (33.0%) and F1Score (39.5%) ones. On the other hand, the IMM generates smaller blobs than the other detectors, so it gets the best Precision results (54.7%), but the worse Recall results (53.7%). For these reasons, all the trackers initialized with appearance-motion combinations get worse Recall results than their appearance initialized versions. The trackers initialized with the ISM, TUD, ISM+IMM, TUD+IMM and DTDP+IMM person detection modules get more balanced Precision and Recall results, but the ISM and ISM+IMM variations get good F1Sore results and also the lower DScore results. Finally, in all cases, the combination of appearance and motion introduces a slight decrease in the LostRate.

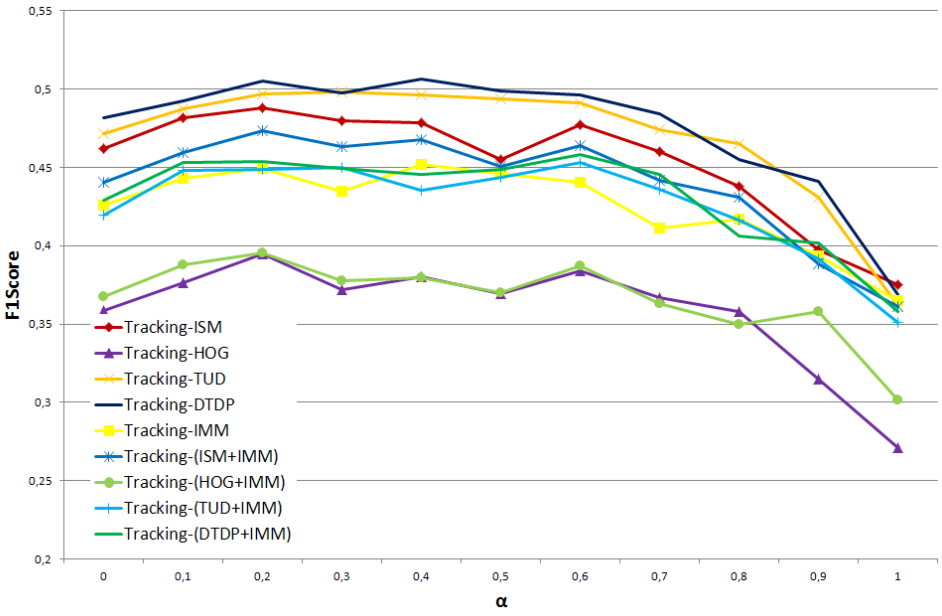


Fig. 6.4. Tracking results according to α update parameter.

Approach	Precision	Recall	F1Score	DScore	LostRate
Tracking-ISM	47.4	70.5	48.8	0.38	15.7
Tracking-HOG	33.0	78.0	39.5	0.88	13.5
Tracking-TUD	49.2	65.6	49.7	0.44	16.0
Tracking-DTDP	48.6	72.6	50.5	0.42	14.7
Tracking-IMM	54.7	53.7	44.9	0.32	18.7
Tracking-(ISM+IMM)	49.7	66.0	47.3	0.40	13.5
Tracking-(HOG+IMM)	38.7	71.0	39.5	0.69	13.1
Tracking-(TUD+IMM)	47.4	63.0	44.8	0.48	14.4
Tracking-(DTDP+IMM)	48.0	68.4	45.3	0.44	14.1

Table 6.3: Tracking results. Different Tracking-Detector initializations results.

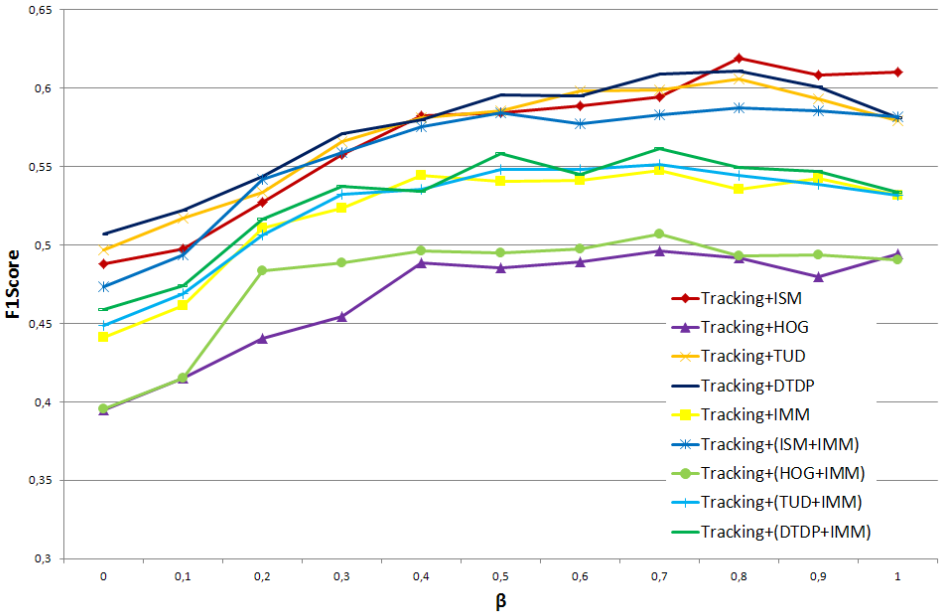


Fig. 6.5. Collaborative system tracking results according to β update parameter.

6.4.4 Collaborative system results

The previous sections describe the experimental results with the two first system configurations: detection and tracking. This section discusses the results obtained by updating each process with the information provided by the other process, that is, the experimental results with the collaborative detection-tracking system configuration. The tracking update approach has been evaluated with different variations of the update parameters ($0 \leq \beta \leq 1$), obtaining a common pattern (see Figure 6.5). To discuss the collaborative system results, the update parameter of the proposed method has been fixed to $\beta=0.8$ (with $\alpha=0.2$ as previously fixed).

6.4.4.1 People detection results

This section describes the results obtained in the detection task using the collaborative detection-tracking system configuration. It means that the detection process undergoes an update using the information provided by the tracker instances (i.e. the tracking process). Following the same evaluation scheme as in section 6.4.2, we re-evaluate the people detection task. Table 6.4 summarizes the obtained detection results. As mentioned previously, in general, all algorithms had high Precision values (93∼96%) and low Recall values (10∼26%) (see Table 6.2). We try to correct the people detection "unstable" behavior over time with the tracking information. The people detection probability prediction and update allows us to stabilize the detection over time and to eliminate false positive detections quickly. For this reason, there is a significant

Approach	Precision	%Δ	Recall	%Δ	F1Score	%Δ
ISM+Tracking	94.1	-0.6	24.3	+47.3	36.6	+34.6
HOG+Tracking	94.8	+1.5	19.0	+23.4	28.6	+13.0
TUD+Tracking	94.4	-0.7	19.6	+81.1	30.0	+57.1
DTDP+Tracking	94.9	-0.2	23.1	+18.5	34.2	+11.4
IMM+Tracking	94.4	-0.7	19.8	+63.6	30.4	+43.4
(ISM+IMM)+Tracking	94.8	+1.0	28.8	+32.7	42.7	+23.4
(HOG+IMM)+Tracking	95.8	-0.6	29.4	+36.7	43.3	+25.5
(TUD+IMM)+Tracking	93.0	-1.5	25.2	+48.2	37.8	+32.6
(DTDP+IMM)+Tracking	94.4	-2.1	28.1	+6.8	42.2	+7.7

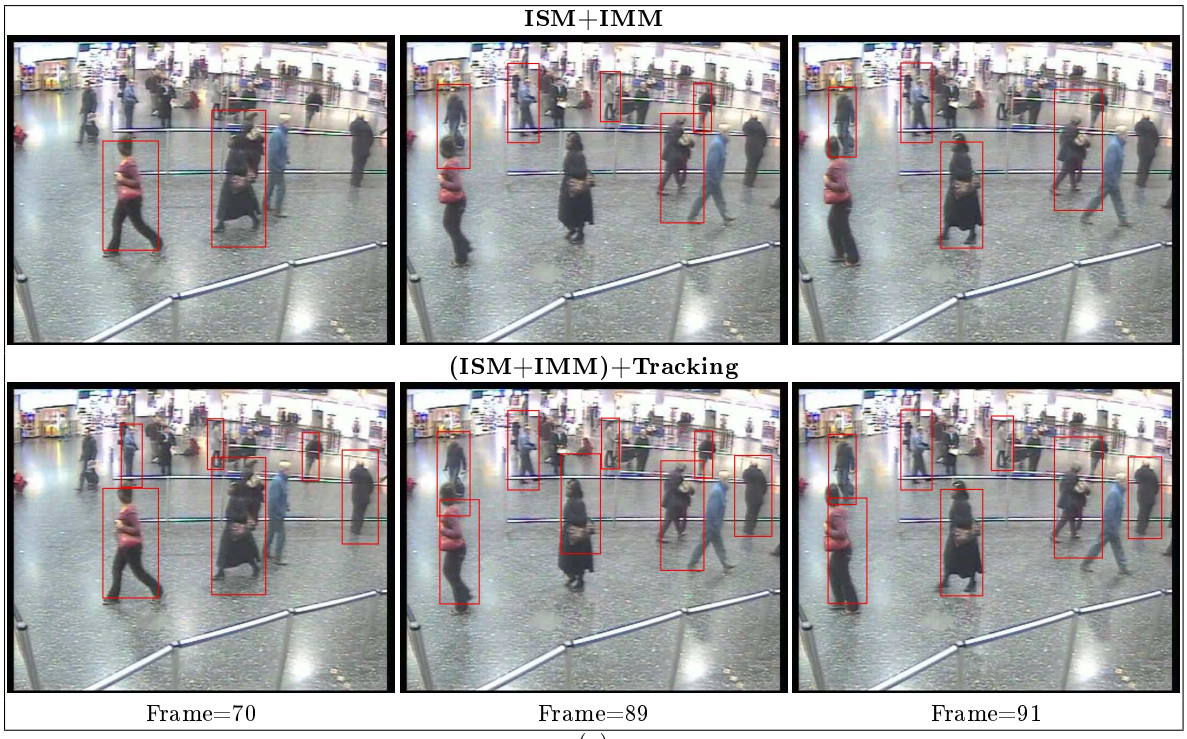
Table 6.4: Collaborative system people detection results. Percentage increase (%Δ) calculated with respect to detectors without people detection update process.

improvement in terms of Recall (6∼81%) and F1Score (7∼57%) without a significant Precision variation. As in the detection without tracking information, the appearance-motion combinations are clearly better than their single versions and the ISM+IMM+Tracking, HOG+IMM+Tracking and DTDP+IMM+Tracking detectors get similar results, whilst the TUD+IMM detector is clearly worse.

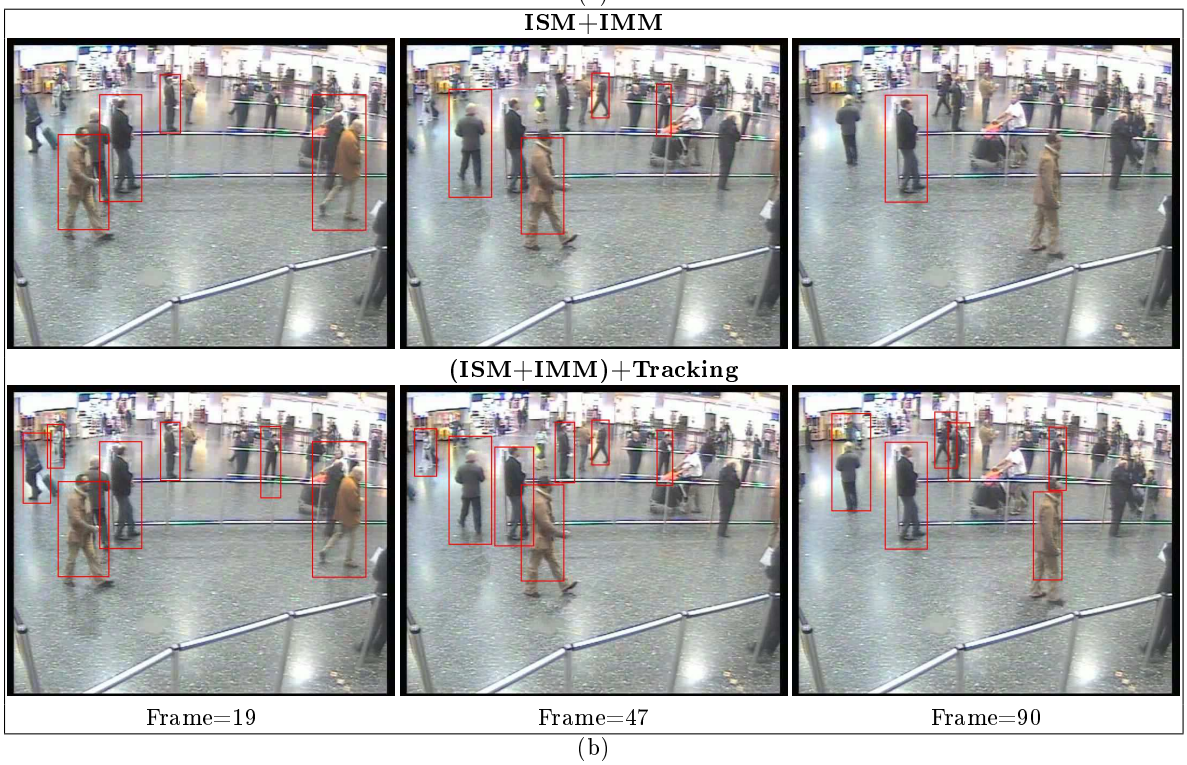
Figure 6.6 shows two visual examples of the difference in people detection performance with and without using our collaborative approach. In order to show the visual results, the ISM+IMM detector (one of the detectors with the best results) has been chosen. The people detection instability over time can be easily observed, as the ISM+IMM detector presents an intermittent performance: there are multiple cases of people detected in one frame and not detected in consecutive frames, even with a minimal variation of appearance (in Figure 6.6(a) there is one example of a non-detected woman on the center of the image in the frame 89 and the same woman correctly detected in frames 70 and 91). With the tracking information (ISM+IMM+Tracking), we are able to extrapolate in time these intermittent detections (see Figures 6.6(a) and (b)) and dismiss quickly this extrapolation in time according to a fast tracker confidence decrease in the cases of people going out of the scene or temporally occluded people (see Figure 6.6(b)).

6.4.4.2 Tracking results

This section describes the results obtained in the tracking task using the collaborative detection-tracking system configuration. It means that the tracking process undergoes an update using the information provided by the people detection. Following the same evaluation scheme as in section 6.4.3, we re-evaluate the tracking task. Table 6.5 summarizes the obtained tracking results. We can see how the use of any people detector significantly improves the tracking performance, as it allows us to correct the possible tracker drifting over time. Percentage increases between 22∼27%



(a)



(b)

Fig. 6.6. People detection results: people detection (ISM+IMM) vs. collaborative system people detection ((ISM+IMM)+Tracking). Test sequences (referring to PDds numbering): (a) 23 and (b) 46.

Approach	Precision	%Δ	Recall	%Δ	F1Score	%Δ	DScore	Δ	LostRate	%Δ
Tracking+ISM	59.8	+26.2	80.3	+13.9	61.9	+26.8	0.21	-0.17	15.4	-1.9
Tracking+HOG	41.7	+26.4	83.1	+6.5	49.2	+24.6	0.68	-0.20	14.2	+5.2
Tracking+TUD	61.4	+24.8	72.7	+10.8	60.6	+21.9	0.41	-0.03	16.1	+0.6
Tracking+DTDP	59.4	+22.2	81.7	+12.5	61.1	+21.0	0.35	-0.07	15.1	+2.7
Tracking+IMM	66.8	+22.1	58.9	+9.7	53.6	+19.4	0.32	0	18.5	-1.1
Tracking+(ISM+IMM)	62.0	+24.7	73.8	+11.8	58.7	+24.1	0.24	-0.16	13.2	-2.2
Tracking+(HOG+IMM)	49.0	+26.6	75.2	+5.9	49.3	+24.8	0.51	-0.18	14.1	+7.6
Tracking+(TUD+IMM)	59.5	+25.5	65.6	+4.1	54.5	+21.7	0.44	-0.04	13.0	-9.7
Tracking+(DTDP+IMM)	61.4	+27.9	70.1	+2.5	55.0	+21.4	0.40	-0.04	13.7	-2.83

Table 6.5: Collaborative system tracking results. Increase (Δ) and percentage increase (%Δ) calculated with respect to trackers without tracking update process.

in Precision, between 2∼13% in Recall and between 19∼26% in F1Score are achieved. Moreover, there is a general improvement on the DScore results. As in the tracking without people detection information, the HOG algorithm gets the best Recall results (83.1%), but the worse Precision (41.7%) and F1Score (49.2%) ones. On the other hand, the IMM algorithm gets the best Precision results (66.8%) and the worse Recall results (58.9%), which is still influencing on the appearance-motion versions Recall results. Again, the trackers with the ISM, TUD, DTDP, ISM+IMM, TUD+IMM or DTDP+IMM algorithms get more balanced Precision and Recall results, whilst the ISM and ISM+IMM variations still get good F1sore results and also the lower DScore results. And, in all cases, the combination of appearance and motion introduces a slight decrease in the LostRate. In general, the collaborative detection-tracking system maintains or introduces a slight percentage decrease in the LostRate, but in the cases of the use of HOG and DTDP detectors, it introduces a small percentage increase (between 5∼7%) and (2%) respectively.

Figure 6.7 shows two visual examples of the difference in tracking performance with and without using our collaborative approach. In order to show the visual results, the Tracking-(ISM+IMM) tracker (one of the trackers with the best results) has been chosen. The tracking difficulties can be easily observed, as the Tracking-(ISM+IMM) tracker presents drifting problems: there are multiple cases of drifting examples (see one example in Figure 6.7(a), the man on the center of the image and a black blob in the frame 95, and two examples in Figure 6.7(b), the man on the center of the image and a green blob in the frame 97 and the woman on the right side of the image and a magenta blob in the frame 44). With the people detection information (Tracking+(ISM+IMM)), we are able to correct these cases of drifting and, therefore, improve the global tracking performance.

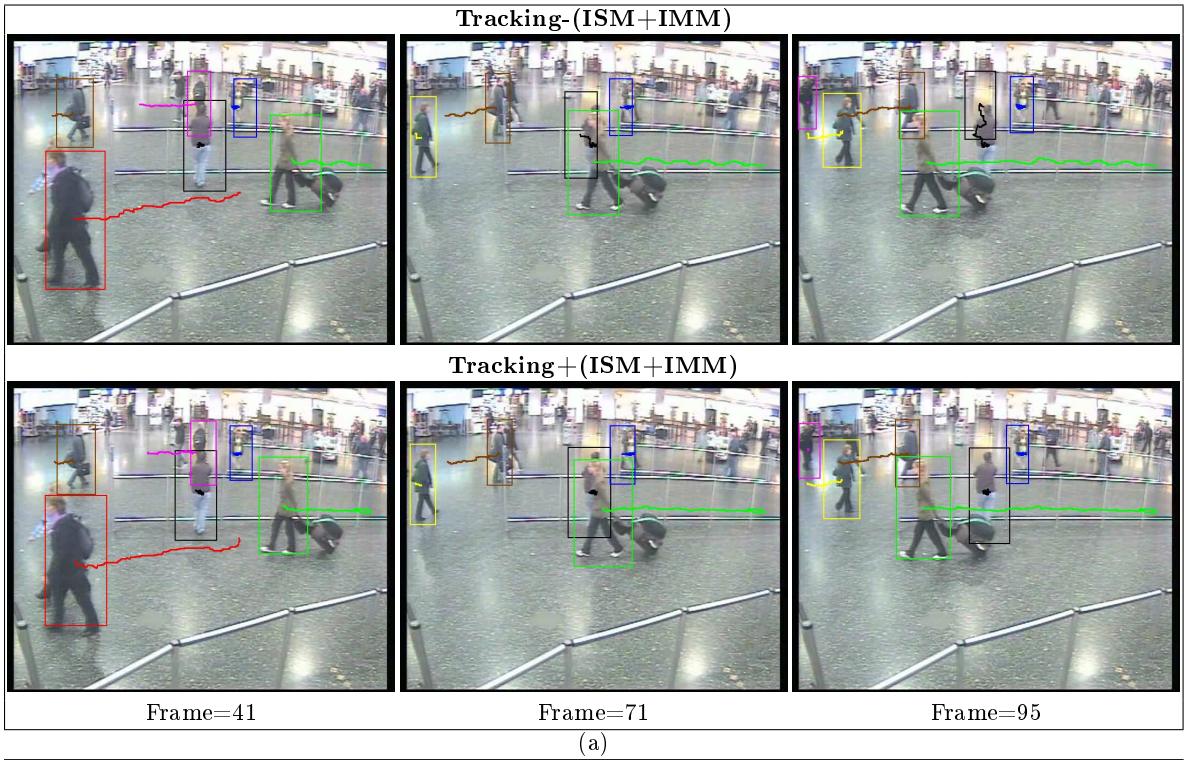


Fig. 6.7. Tracking results: tracking (Tracking-(ISM+IMM)) vs. collaborative system tracking (Tracking+(ISM+IMM)). Test sequences (referring to PDds numbering): (a) 14 and (b) 54. For visualization purposes, the shown trajectories are computed by a median filter (order N=5). However, only the blobs are used for evaluation.

6.4.4.3 Collaborative systems discussion

A straight comparison with other systems from the state of the art that combine detection and tracking is quite difficult, not only because of the difficulty to replicate faithfully any approach, but also because of the variety of evaluation methodology and experimental dataset (see Table 6.1): some approaches do not perform a comparison of the improvement introduced in one or both tasks, or directly do not evaluate one of the tasks.

Our system has been designed with the aim of improving both tasks simultaneously and being able to evaluate each task independently and their combination. In return, most approaches from the state of the art that combine detection and tracking are designed mainly with the aim of improving tracking results [Leibe et al., 2007; Andriluka et al., 2008; Ess et al., 2009; Stalder et al., 2010; Yu et al., 2011; Wu and Nevatia, 2007; Ren, 2008; Giebel et al., 2004; Okuma et al., 2004; Avidan, 2007; Li et al., 2008; Breitenstein et al., 2010]. There are some approaches that also try to improve explicitly the detector [Avidan, 2007; Li et al., 2008; Breitenstein et al., 2010], but they only make use of the tracker trajectory and only with the main objective of improving the tracking. We propose a people detection update which includes the trajectory and the tracker confidence level, allowing us not only to extrapolate missing detections in time, but also to dismiss quickly this extrapolation in time according to a fast tracker confidence decrease (person going out of the scene or occlusions). Moreover, unlike the approaches discussed from the state of the art, our system is focused on improving both tasks simultaneously and not only the tracking, so an exhaustive evaluation has been performed (dataset B), including both tasks independently and their combination.

6.4.5 Computational cost

According to the computational cost, each detector results have been obtained with the available code, implemented with different tools and programming languages, so a fair comparison is not possible. However, assuming that the detection and tracking processes run in parallel and being almost insignificant the computational cost of the information exchange process, the final combination approach computational cost will be established by the detection approach or the tracking approach.

According to the original implementations, the ISM approach [Leibe et al., 2008] computational cost is between 4-7 seconds per frame with 320x240 images, the HOG approach [Dalal and Triggs, 2005] computational cost is around 1 second per frame with 352x288 images (there is a faster implementation in OpenCV that runs around 0.1 seconds per frame or around 10 frames per second -see section 4.3-), the TUD approach [Andriluka et al., 2009] computational cost is several orders of magnitude greater than the other approaches and the DTDP approach [Felzenszwalb et al., 2010] computational cost is around 2 seconds per frame with 640x480 images. Finally, the tracking approach [Nummiaro et al., 2003] computational cost is around 0.05

seconds per frame or around 20 frames per second with 360x288 images. Therefore, the computational cost will be established by the detection process and any detection (ISM, HOG; TUD, DTDP and IMM) and tracking combinations will be far from the real time operation.

6.5 Summary and conclusions

In this chapter, a collaborative people detection and tracking system is proposed. It integrates the people detection and tracking information into a single system and improves both tasks simultaneously. We have analyzed the three system configurations in order to evaluate the improvement introduced by the mutual information exchange. Experiments have been conducted on challenging sequences extracted from the PDds dataset created with TRECVID sequences (highly crowded scenes, severely cluttered background and people at different scales), highlighting the problems that these complex scenarios entail in the state of the art of people detection and tracking. The experiments on the proposed dataset show the utility of the collaborative system, especially in complex scenarios, getting better results than the state of the art for each task independently. The detection and tracking modules can be replaced by others without great difficulty thanks to the modular design of the system that allows a collaborative or independent performance, the generic format of the information to be exchanged (blobs and detection/tracking confidence) and the easily compatible information exchange mechanism (simple and consistent process updates). The use of different modules will vary the overall performance of the system, but the combination of both sources of information will always be useful for improving the system (except in the ideal case of perfect detection and perfect tracking).

With respect to people detection, firstly, we have used a people detector based on the combination of appearance and motion information. We have evaluated different appearance-motion combinations of people detectors from the state of the art and it is clear that human motion provides useful information for people detection and independent from appearance information. Secondly, a people detection prediction or update scheme using the tracking information about our collaborative system has been proposed and all the different people detector variations have been re-evaluated. The experimental results show that the use of tracking information stabilizes the people detection over time, so there is a significant improvement mainly in terms of Recall and F1Score.

With respect to tracking, in a first place, an adaptive particle filter tracker based on color distributions with different people detection initializations has been evaluated. All trackers follow a similar pattern, but it is shown clearly that the initialization has a great influence on the global tracker performance. Secondly, all the tracker variations have been re-evaluated adding the people detection information about our collaborative system. The experimental results show that the use of people detection information corrects the position, dimension and color distribution of

the trackers over time, so there is a significant improvement mainly in terms of Precision and F1Score.

As already commented, the detection and tracking modules can be replaced by others from the state of the art without great difficulty. The use of different modules will vary the overall performance of the system, but the combination of both sources of information, in principle, will be useful for improving the system. In the following chapters, we are focused on the detection module, proposing post-processing subtasks in order to improve the detection performance in typical video surveillance environments.

Chapter 7

People detection using people-background segmentation confidence

7.1 Introduction[1]

Again, people detection is one of the most challenging problems in computer vision. People detection approaches from the state of the art obtain satisfactory results in low and medium complexity scenarios, but these results are considerably reduced in more complex and realistic scenarios (see chapter 4). In order to achieve a more reliable performance in complex scenarios, we have proposed a new people detection approach based on motion and their combination with appearance information (see chapter 5) and we have also proposed the integration of this appearance and motion information in a detection and tracking system that takes advantage of the tracking information (see chapter 6). In this chapter, we propose a new people detection filtering subtask that reduces the number of false positive detections and, therefore, improves the global detection results.

A people-background segmentation is a two-class segmentation ensuring that no people or body parts are appearing in the background class. This segmentation is desirable for many computer vision applications, such as robotics and driver assistance systems. This type of segmentation is useful not only as a people detection preprocessing or post-processing step, but also for other video analysis processes such as tracking and people density estimation. While the focus of person detection approaches is to obtain a high detection performance and to reduce false positive detections, we aim at determining the areas without people in the scene by giving

[1] This chapter is based on the publication *"A. García-Martín, A. Cavallaro, J. M. Martínez. People-background segmentation with unequal error cost. In Proc. of the IEEE International Conference on Image Processing, 2012"*

a higher penalty to pixels representing a person, but that have been incorrectly classified as background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with bias on people.

Despite the fact that the state of the art in people detection includes several solutions working in specific and constrained scenarios, every people detection approach presents limitations and drawbacks. In this chapter, we address one of the main problems of people detection in video sequences: every people detector from the state of the art must maintain a balance between the number of false detections and the number of missing pedestrians. This compromise limits the global detection results. In order to reduce or relax this limitation and improve the detection results, we propose to use the people-background segmentation as a filtering stage in people detection.

The main objective of this chapter is to present a new people detection filtering subtask based on the people-background segmentation. People-background segmentation gives us information about where there are not people in the scene. We can use this information to eliminate, or at least reduce, the number of false positives and, therefore, improve the global detection results.

In this chapter, we will firstly make a brief introduction to the related literature in section 7.2. The proposed people-background approach is described in section 7.3. Then, the proposed people detection filtering subtask based on the people-background segmentation is described in section 7.4. After that, section 7.5 describes the experimental results. Finally, section 7.6 summarizes the chapter with some conclusions.

7.2 Related work

Traditionally, the typical additional preprocessing subtasks in people detection are not oriented to one specific processing task, i.e., they are oriented to enhance, adapt or reduce the video information before being analyzed, for example: camera motion compensation, camera calibration, noise removal, etc. In return, the typical additional post-processing subtasks in people detection are applied over the detection outcome and are oriented to filter or verify the final detections using any additional information source. The most typical ones are those based on tracking information [Ess et al., 2009], which study the detections evolution over time. Other approaches use some kind of scene or contextual restriction [Gerónimo et al., 2010] (spatial, people size, symmetry, etc).

People-background segmentation consists of a two-class segmentation with unequal error cost between classes in order to ensure that no body parts are classified as background. While the focus of person detection approaches is to obtain a high detection performance and to reduce false positive detections, we aim at determining the areas without people in the scene by giving a higher penalty to pixels representing a person, but that have been incorrectly classified as

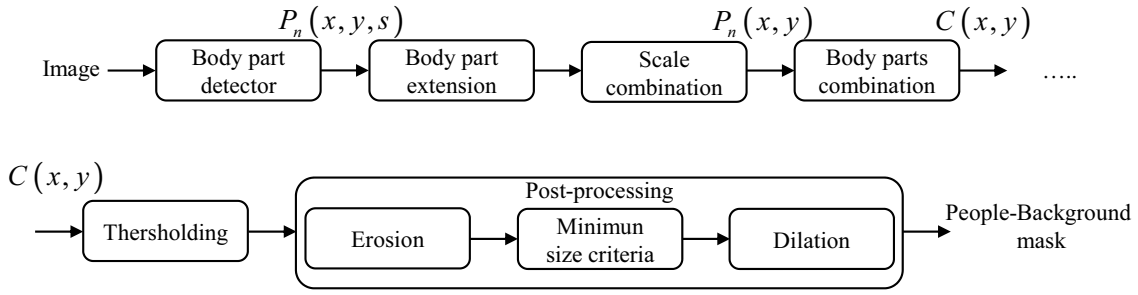


Fig. 7.1. Block diagram of the proposed people-background segmentation approach.

background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with bias on people.

In this work, we propose a people detection approach that enhances people detection results making use of the information about where there are not people in the scene obtained with the people-background segmentation. The proposed filtering approach has been implemented as a post-processing, but it can be used as either a preprocessing or post-processing stage. Experiments have been performed on an extensive dataset with different approaches from the state of the art and show the benefits achieved using the people-background segmentation information.

7.3 People-background segmentation with unequal error cost

The proposed people-background segmentation method is based on [Felzenszwalb et al., 2010] for detecting body parts and extends this representation by appropriately grouping them. Then, we fuse detection confidence maps according to regions that are expected to be covered by the body parts. The corresponding background segmentation mask is finally generated after binarization and post-processing (Figure 7.1).

Starting from the body-part representation introduced in [Felzenszwalb et al., 2010], in this section we define five methods: an independent body parts approach, IBP; a dependent body parts approach, DBP; their extended versions, IEBP and DEBP, respectively; and the post-processed version of DEBP, which we will refer to as DEBP-P.

Let us consider the part-based multi-scale detector (Figure 7.2(a)), where $P_n(x,y,s)$ represents the confidence at pixel position (x,y) for body part n $(n = 1,...,N)$ associated to scale s $(s = 1,...,S)$. Let also each body part be modeled by a 3-tuple $(F_n,\ v_{n,0},\ d_n)$ [Felzenszwalb et al., 2010], where F_n is the HOG filter response (detection confidence) [Dalal and Triggs, 2005] for part n; $v_{n,0}$ is a two-dimensional vector defining the relative position of part n with respect to the anchor position (x_0, y_0) of the whole body; and d_n is a four-dimensional vector specifying coefficients of a quadratic function defining the cost for each possible placement of the part relative to the anchor position. The confidence score for part n at scale s is given as

$$P_n(x, y, s) = F_n(x, y, s) - \langle d_n, \phi(dx_n, dy_n) \rangle \tag{7.1}$$

with

$$(dx_n, dy_n) = (x_n, y_n) - (2(x_0, y_0) + v_{n,0}) \tag{7.2}$$

giving the displacement of part n relative to the anchor and

$$\phi(dx, dy) = \left(dx, dy, dx^2, dy^2\right) \tag{7.3}$$

defining the potential spatial deformation distributions [Felzenszwalb et al., 2010].

We define IBP by using eight ($N = 8$) *independent* body parts I_n, with $n = 1, ..., N$ and specified the anchor position $v_{n,n}$ relative to the body part n instead of the root position (Figure 7.2(b)). To improve the detection robustness, we then define DBP using M *dependent* body part models D_m, with $m = 1, ..., M$ as combination of independent parts (Figure 7.2(c)). Each D_m is defined by L_m parts, $I_1, ..., I_{L_m}$, where I_{l_m} is one of the independent parts with its anchor position $v_{l,m}$ relative to the *corresponding* dependent body part D_m. In order to exploit the correlation between body parts, we have chosen $M = 4$ dependent body parts: head and shoulders, trunk, legs and full body. Moreover, in order to recover undetected dependent body parts or normalize the detection confidence between dependent body parts already detected, we propose to extend the dependent body parts definition and reuse the information from other dependent body parts. Each dependent body part D'_m is given by the maximum between the original dependent body part D_m and the average of the other dependent body parts, all of them relative to the same D_m.

If we assume that there are at least two visible dependent body parts for each person, we are able to recover or normalize body parts by averaging the remaining parts and, in turn, we avoid the reproduction of those isolated dependent body parts incorrectly detected:

$$D'_m(x, y, s) = max\left(D_m(x, y, s), \frac{1}{M-1}\sum_{i \neq m}^{M} D_{i,m}(x, y, s)\right), \tag{7.4}$$

where $D_{i,m}(x, y, s)$ is the body part i with anchor position $v_{i,m}$.

Once we have obtained the dependent or independent body parts responses at each pixel position and scale, the confidence of each body part response is extended to define the methods IEBP and DEBP, respectively (Figure 7.3). IEBP extends each independent body part, whilst DEBP extends each independent body parts combination. Both IEBP and DEBP cover the detected part in the chosen body parts representation as represented by the kernel extensions (yellow shapes) in Figure 7.2(b) and (c) according to the area that it is expected to cover in a frame.

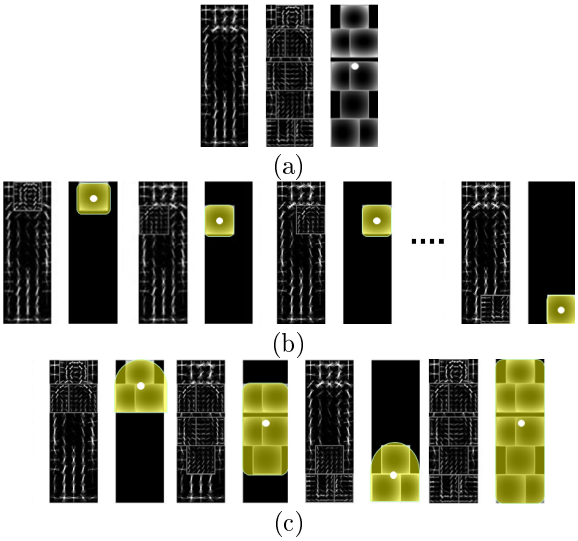


Fig. 7.2. Body parts representations. (a) Multi-part person model from [Felzenszwalb et al., 2010]; (b) IBP model; (c) DBP model. The kernel used in the extensions is shown in yellow.

Once we have obtained all the final body part detection confidence maps $P_n(x, y, s)$, with $n = 1, ..., N$, we select for each position in the frame the maximum confidence level across scales and across parts to generate the fused confidence map $C(x, y)$:

$$C(x, y) = \max_{n=1,...,N} \max_{s=1,...,S} P_n(x, y, s). \tag{7.5}$$

Figure 7.3 shows examples of confidence maps generated on the same frame using the original method [Felzenszwalb et al., 2010], IBP, IEBP, DBP and DEBP.

The final people-background mask is obtained by binarizing $C(x, y)$. Assuming that each person in the scene is visible (i.e. at least two dependent body parts are captured in the frame) or is partially occluded by another person, regions that are smaller than the minimum size of a person are eliminated. The minimum size is defined by the person model scale in [Felzenszwalb et al., 2010]. The resulting mask undergoes an erosion with a disc the size of the smallest body part to detect in the minimum size of a person, followed by connected components analysis to remove regions that are smaller than the minimum size of a person. Finally a dilation operation with a disc the size of the smallest body part to detect in the maximum size of a person is performed to generate the final mask. We will refer to this overall method as DEBP-P. Sample results are shown in Figure 7.4.

The experimental evaluation details of the proposed people-background segmentation approach are described in appendix B. This chapter is focused on the people detection filtering subtask using the people-background segmentation information.

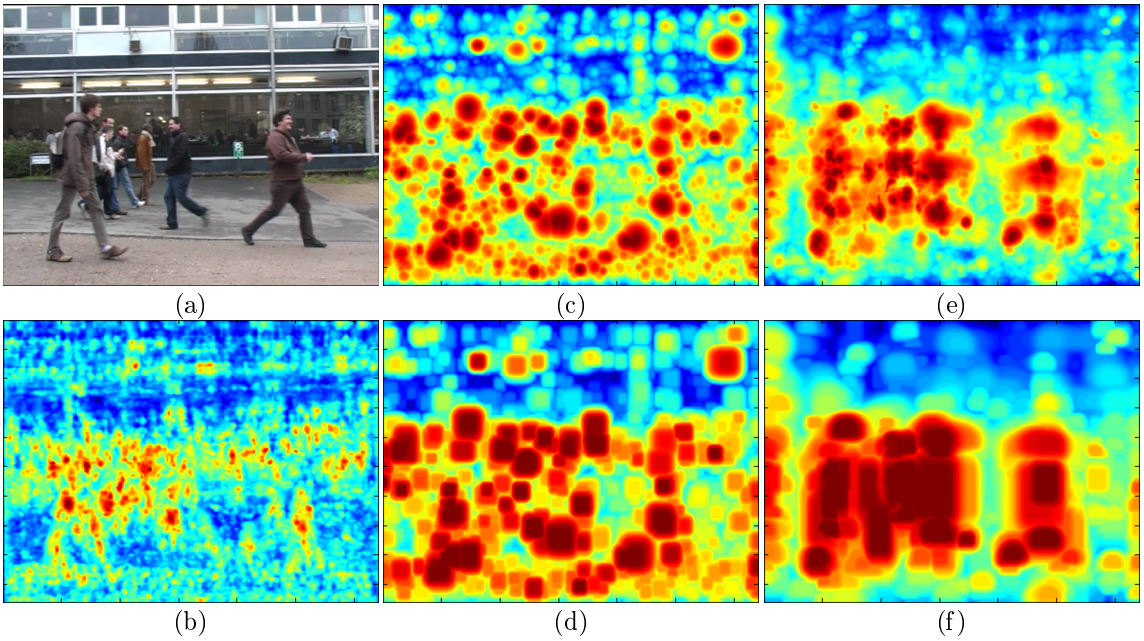


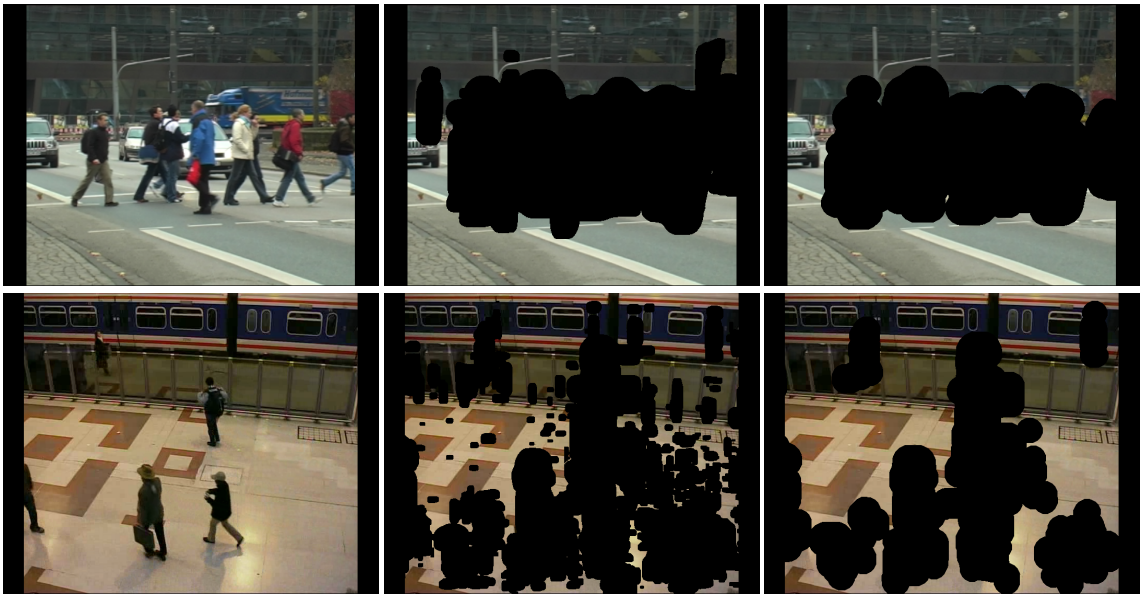
Fig. 7.3. Confidence maps for a sample frame (a) generated with: (b) the original method [Felzenszwalb et al., 2010]; (c) IBP; (d) IEBP; (e) DBP; and (f) DEBP.

7.4 People detection using people-background segmentation

As already mentioned, every people detector from the state of the art must keep a balance between Precision and Recall rates. For this reason, the global detection performance is mainly limited by the number of possible false detections. Our main idea consists of reducing or relaxing this limitation using the people-background segmentation.

In this work, we propose a people detection system that includes a post-processing stage using the people-background segmentation information (see Figure 7.5). Firstly, people detections are obtained using any people detector from the state of the art and the people-background segmentation is obtained as described in the previous section 7.3. Then, both information sources are combined with the aim of eliminating or reducing the number of false detections, but maintaining, as much as possible, the number of positive detections. Figure 7.6 shows one experimental example where it is shown that depending on the selected threshold the number of true positives are maintained reducing false positives (straight line) or reduced, but reducing more the number of false positives (dotted line). The combination is made with the detections (bounding box and people detection confidence) and with the people-background confidence map (DEBP confidence map -see previous section 7.3-) or the binarized and post-processed segmentation mask (DEBP-P segmentation mask -see previous section 7.3-).

In general, any people detection outcome always consists of a list of N detections in each frame t. Each detection n ($n = 1, ..., N$) is represented by its position (x, y) and dimensions (w, h), i.e.,



(a) (b) (c)

Fig. 7.4. Examples of results: (a) sample image; (b) DEBP result; (c) DEBP-P result.

bounding box (blob) $B_n(x, y, w, h)$ and a People-detection Confidence PC_n $(0 \leq PC_n \leq 1)$. In order to process every detection, it has been defined a Segmentation Confidence associated with every detection SC_n $(0 \leq SC_n \leq 1)$. This associated confidence is the averaged segmentation confidence over the corresponding blob (see equation 7.6).

In the case of the DEBP confidence map $C(x, y)$, it is the averaged of the dense confidence values SC_n^C. However, in the case of the DEBP-P segmentation mask $M(x, y)$ (a binarized and post-processed version of the DEBP confidence map), the segmentation confidence corresponds to the percentage of pixels classified as people vs. the number of pixels classified as background SC_n^M:

$$SC_n^{C/M} = \frac{1}{w \cdot h} \sum_{x,y \in B_n} C/M(x, y) \tag{7.6}$$

Figure 7.7 shows SC^C and SC^M examples over a positive and a false detection.

Then the final detections consist of the initial N detections with a new associated confidence based on the combination of the detection and segmentation confidences PSC_n $(0 \leq PSC_n \leq 1)$:

$$PSC_n = PC_n \cdot SC_n \tag{7.7}$$

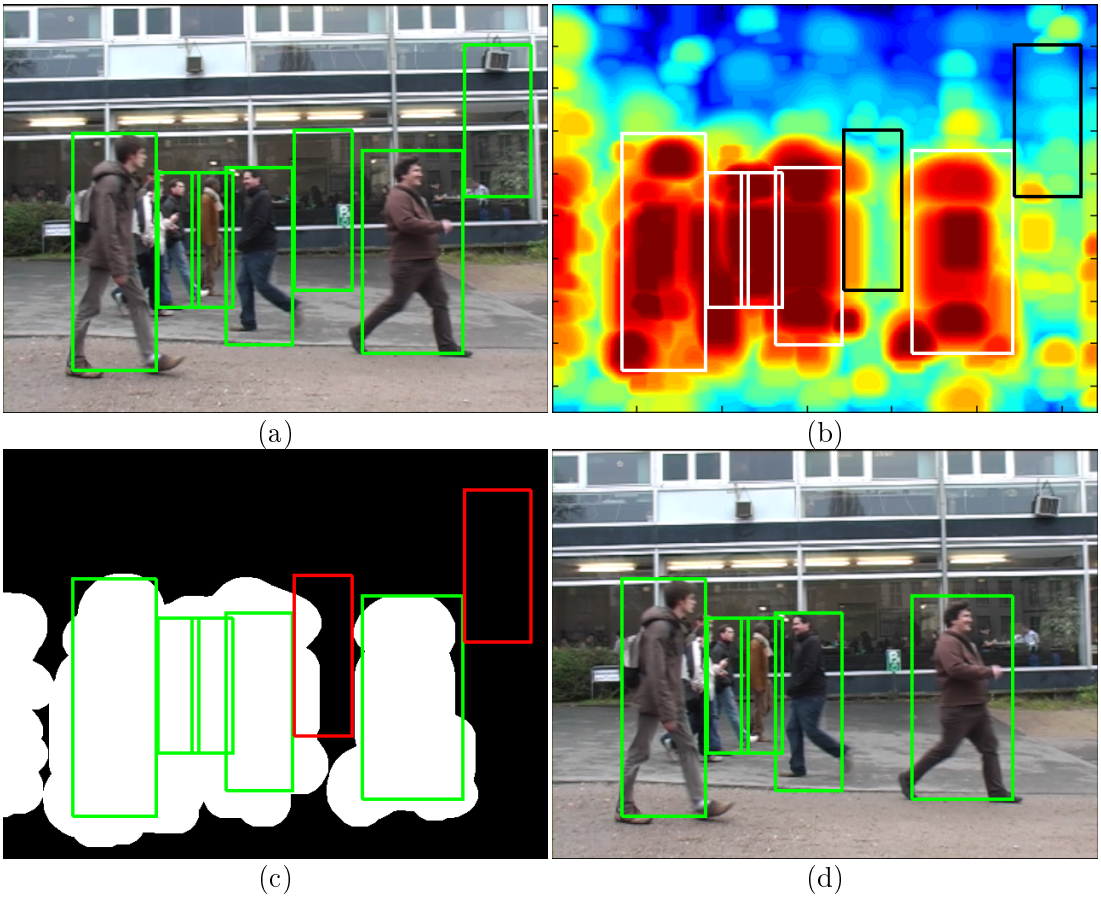


Fig. 7.5. People detection system example: (a) people detections; (b) people detections over the DEBP segmentation confidence map; (c) people detections over the DEBP-P segmentation mask; and (d) final people detections.

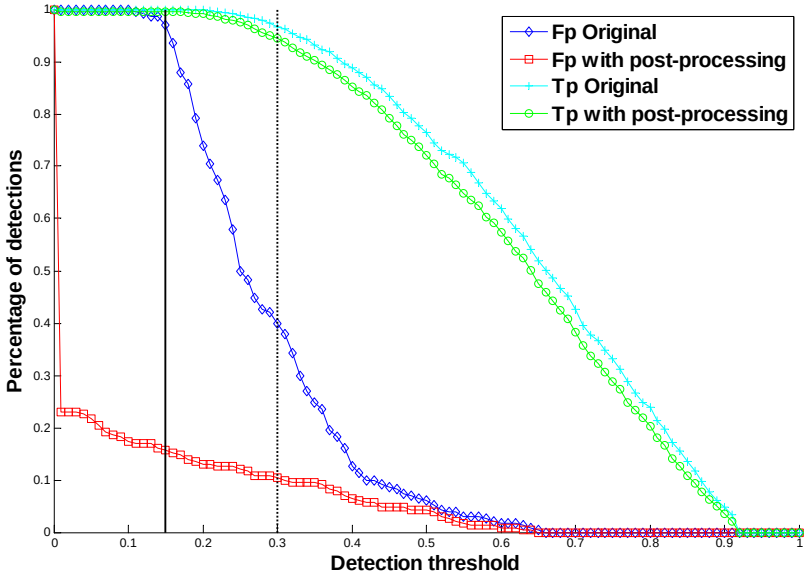
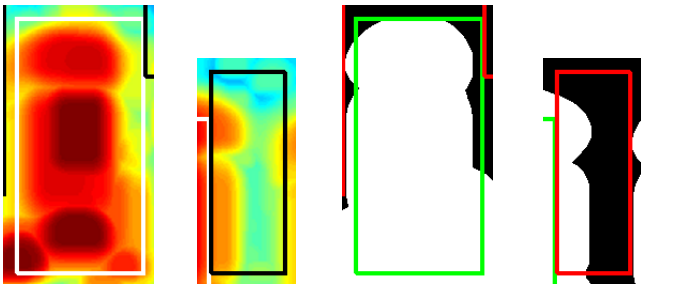


Fig. 7.6. Percentage of false positive (Fp) and true positive (Tp) detections with and without the proposed post-processing. According to the selected threshold, the number of true positives are maintained reducing false positives a 81% (straight line or 0.15 threshold) or the number of true positives are reduced a 3%, but reducing the number of false positives a 29% (dotted line or 0.3 threshold).



$SC^C = 0.87$ (a) $SC^C = 0.50$ (b) $SC^M = 0.93$ (c) $SC^M = 0.38$ (d)

Fig. 7.7. Examples of segmentation confidence $SC^{C/M}$ associated with a positive and a false detection: (a) and (b) using the DEBP confidence map SC^C; (c) and (d) using the DEBP-P segmentation mask SC^M.

7.5 Experimental results

This section describes the experimental setup used in the evaluation of the proposed people detection post-processing stage, the experimental results and the computational cost.

7.5.1 Experimental setup

In order to evaluate our people detection approach, we compare in this section the original performance (see chapter 4) and the post-processed performance after using the people-background information over seven people detection approaches from the state of the art: Edge (see chapter 4), Fusion [Fernández-Carbajales et al., 2008], HOG [Dalal and Triggs, 2005], ISM [Leibe et al., 2005], TUD [Andriluka et al., 2009], DTDP [Felzenszwalb et al., 2010] and IMM (see chapter 5). There is a brief description of the different people detection approaches used from the state of the art in appendix A.

As in the chapter 4, focused on the idea of evaluating the performance of the proposed approach in different typical video surveillance environments, the proposed approach has been evaluated in both evaluation datasets (A and B) described in the performance evaluation methodology (see section 3.3). The dataset A allows us to evaluate the different approaches at every complexity level (C1,..,C5), while the dataset B allows us to evaluate more thoroughly the highest complexity category (C5). The different detection approaches experimental results have been obtained using the available code and binaries. In the case of the people-background segmentation, it has been obtained using the original code and the chosen empirical binarization threshold is 0.8 (see Appendix B).

7.5.2 People detection results

7.5.2.1 Evaluation dataset A

Firstly, we evaluate and compare the appearance based people approaches at every complexity level using the evaluation dataset A. The original people detection results have been already discussed in chapter 4.

Tables 7.1 and 7.2 show the people detection results using the DEBP confidence map and the DEBP-P segmentation mask respectively. The use of the people-background segmentation allows us to reduce the number of false detections and, therefore, in almost all the cases we improve the global detection results. The improvements obtained with the DEBP-P segmentation mask (average improvement of 5.4%) are significantly better than the ones obtained with the DEBP confidence map (average improvement of 3.8%) with the inconveniences of binarization (defining a segmentation threshold and computing some post-processing, e.g., erosion and dilatation -see section 7.3-).

	Edge	%Δ	Fusion	%Δ	HOG	%Δ	ISM	%Δ	TUD	%Δ	DTDP	%Δ	Total	%ΔTotal
C1	0.99	+1.0	0.88	+12.8	0.91	-1.1	0.98	+3.2	0.96	+3.2	0.96	+0.0	0.95	+3.2
C2	0.95	+2.2	0.83	+2.5	0.86	+0.0	0.93	+2.2	0.91	+3.4	0.92	+0.0	0.90	+1.7
C3	0.90	+5.9	0.68	+13.3	0.79	+6.8	0.87	+8.8	0.83	+10.7	0.85	+4.9	0.82	+8.4
C4	0.89	+0.0	0.72	+4.3	0.83	+1.2	0.87	+3.6	0.85	+1.2	0.87	+1.2	0.84	+1.9
C5	0.73	+4.3	0.52	+8.3	0.73	+2.8	0.74	+4.2	0.70	+4.5	0.74	+0.0	0.69	+4.0
Total	0.89	-	0.73	-	0.82	-	0.88	-	0.85	-	0.87	-	0.84	-
%ΔTotal	-	+2.7	-	+8.3	-	+1.9	-	+4.4	-	+4.6	-	+1.2	-	+3.8

Table 7.1: People detection performance using the DEBP confidence map, in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A. Percentage increase (%Δ) calculated with respect to original performance (see section 4.4.2.1).

	Edge	%Δ	Fusion	%Δ	HOG	%Δ	ISM	%Δ	TUD	%Δ	DTDP	%Δ	Total	%ΔTotal
C1	0.99	+1.0	0.96	+23.1	0.91	-1.1	0.98	+3.2	0.97	+4.3	0.96	+0.0	0.96	+5.1
C2	0.95	+2.2	0.84	+3.7	0.86	+0.0	0.94	+3.3	0.91	+3.4	0.92	+0.0	0.90	+2.1
C3	0.92	+8.2	0.71	+18.3	0.78	+5.4	0.89	+11.3	0.87	+16.0	0.87	+7.4	0.84	+11.1
C4	0.90	+1.1	0.72	+4.3	0.84	+2.4	0.88	+4.8	0.86	+2.4	0.87	+1.2	0.85	+2.7
C5	0.74	+5.7	0.54	+12.5	0.73	+2.8	0.75	+5.6	0.72	+7.5	0.75	+1.4	0.71	+5.9
Total	0.90	-	0.75	-	0.82	-	0.89	-	0.87	-	0.87	-	0.85	-
%ΔTotal	-	+3.6	-	+12.4	-	+1.9	-	+5.6	-	+6.7	-	+2.0		+5.4

Table 7.2: People detection performance using the DEBP-P segmentation mask, in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A. Percentage increase (%Δ) calculated with respect to original performance (see section 4.4.2.1).

According to the experimental dataset complexity classification, the results show that in both cases (DEBP and DEBP-P) the highest improvements are obtained in categories C3 (average improvement of 8.4 and 11.1% respectively) and C5 (average improvement of 4.0 and 5.9% respectively). It is due mainly to the background complexity: these two categories (C3 and C5) present medium or high background complexity, being the background complexity one of the main factors that produce false detections. For the same reason, the lowest improvements are obtained in categories C2 (average improvement of 1.7 and 2.1%) and C4 (average improvement of 1.9 and 2.7%) because the complexity of these categories lies on the classification.

According to the chosen people detection approach, the results show that in both cases (DEBP and DEBP-P) the highest improvements are obtained with the Fusion approach (average improvement of 8.3 and 12.4% respectively). It is due mainly to the original algorithm instability against false detections. The other approaches usually present better behavior against false detections and, therefore, get lower improvements. In particular the HOG approach presents even negative results in category C1 because it generates bigger blobs than the other detectors, so the associated segmentation confidences are affected.

7.5.2.2 Evaluation dataset B

In this section, we evaluate more thoroughly the highest complexity category (C5) the appearance based people approaches using the dataset B. The original people detection results have been already discussed in chapter 4. Tables 7.3 and 7.4 show the people detection results using the DEBP confidence map and the DEBP-P segmentation mask respectively.

As in the evaluation of dataset A, in almost all the cases we improve the global detection results: we can see how the improvements obtained with the DEBP-P segmentation mask (average improvement of 3.8%) are significantly better than the ones obtained with the DEBP confidence map (average improvement of 2.3%). In general the improvements obtained with dataset B are smaller than the ones obtained with dataset A. The results are comparable with the results obtained in categories C2 and C4 of dataset A. It is mainly due to that the complexity of dataset B lies not only on background complexity, but also on the classification complexity.

The results show again that in both cases (DEBP and DEBP-P) the highest improvements are obtained with the Fusion approach (average improvement of 6.8 and 9.1% respectively). However, the HOG approach does not present any improvement due mainly to the already commented problem with the dimensions of the blobs.

7.5.2.3 Evaluation dataset B with motion

In this section, we evaluate again the dataset B with the appearance based people approaches, but in this case including the people detector based on motion IMM and all the appearance and motion combinations (Edge+IMM, Fusion+IMM, HOG+IMM, ISM+IMM, TUD+IMM and

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.60	0.47	0.66	0.69	0.58	0.69	0.62
C5 (%Δ)	+1.7	+6.8	+0.0	+0.0	+3.6	+1.5	+2.3

Table 7.3: People detection performance using the DEBP confidence map, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (%Δ) calculated with respect to original performance (see section 4.4.2.2).

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.61	0.48	0.66	0.70	0.60	0.69	0.62
C5 (%Δ)	+3.4	+9.1	+0.0	+1.4	+7.1	+1.5	+3.8

Table 7.4: People detection performance using the DEBP-P segmentation mask, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (%Δ) calculated with respect to original performance (see section 4.4.2.2).

DTDP+IMM). Firstly, we evaluate the original performance and, then, the results obtained using the proposed post-processing subtask.

In order to evaluate the original IMM detector performance, we need to train the people motion model, therefore, the evaluation dataset B has been divided in training and test. To be homogeneous, the appearance based detectors approaches also have been evaluated on the same video sequences, the test dataset. As in the experiments in chapter 5, the training dataset is composed of 25 sequences and the test dataset is composed of the other 36 sequences. Table 7.5 shows the results in terms of AUC-PR of test dataset.

The results show that the IMM approach gets good results in complex and realistic scenarios and comparable to the other approaches from state of the art. The IMM is based only on motion, so it is only able to detect moving people. For this reason, the IMM approach in general is able to get high precision rates, but low recall rates. Even so, in environments as complex as these ones, the use of motion information obtains results close to the use of appearance information. The combination of appearance and motion information (Edge+IMM, Fusion+IMM, HOG+IMM, ISM+IMM, TUD+IMM and DTDP+IMM) improves the global results in all the cases (average improvement of 6.1%). Thus, it is clear that human motion provides useful information for people detection and independent from appearance information.

Tables 7.6 and 7.7 show the people detection results using the DEBP confidence map and the DEBP-P segmentation mask respectively. As in the evaluation of dataset A and dataset B without motion, in almost all the cases we improve the global detection results: we can see how the improvements obtained with the single appearance versions with the DEBP-P segmentation mask and with the DEBP confidence map (average improvement of 3.0 and 1.9% respectively) or motion versions with the DEBP-P segmentation mask and with the DEBP confidence map

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.58	0.46	0.66	0.64	0.56	0.67	0.60	0.60

	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.62	0.49	0.68	0.67	0.62	0.70	0.63
C5 (%Δ)	+6.9	+6.5	+3.0	+4.7	+10.7	+4.5	+6.1

Table 7.5: Area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to single appearance versions.

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.59	0.48	0.65	0.66	0.58	0.68	0.61	0.62
C5 (%Δ)	+1.7	+4.3	-1.5	+3.1	+3.6	+1.5	+1.9	+3.3

	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.63	0.51	0.68	0.69	0.64	0.71	0.64
C5 (%Δ)	+1.6	+4.1	+0.0	+3.0	+3.2	+1.4	+2.2

Table 7.6: People detection performance using the DEBP confidence map, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original performance.

(average improvement of 2.8 and 2.2% respectively) are quite similar than the obtained with the dataset B without motion (see previous section 7.5.2.2). However, the results show how the use of motion in addition to the use of the proposed post-processing obtains the best final results with the DEBP-P segmentation mask and with the DEBP confidence map (AUC-PR Total average of 65 or 64% respectively).

	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.60	0.48	0.66	0.66	0.60	0.68	0.61	0.62
C5 (%Δ)	+3.4	+4.3	+0.0	+3.1	+7.1	+1.5	+3.0	+3.0

	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.64	0.51	0.68	0.69	0.65	0.71	0.65
C5 (%Δ)	+3.2	+4.1	+0.0	+3.0	+4.8	+1.4	+2.8

Table 7.7: People detection performance using the DEBP-P segmentation mask, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original performance.

7.5.3 Computational cost

According to the computational cost, being almost insignificant the computational cost of the combination of detection and segmentation confidences, we only introduce the additional computational cost of the people-background segmentation. The people-background segmentation is based on the DTDP detector [Felzenszwalb et al., 2010] and has a comparable computational cost. The DEBP-P has an average computational cost increase of 1 second per frame with respect to original approach DTDP (see Appendix B). The proposed approach has been implemented as a post-processing stage in a people detection system, but it could be also applied as a preprocessing step with similar detection results and allowing a computational cost reduction of the subsequent people detector approach and, therefore, the global computational cost.

7.6 Summary and conclusions

We have presented a new people detection filtering subtask based on the people-background segmentation. People-background segmentation gives us information about where there are not people and, therefore, the possibility of filtering or reducing the false positive detections in those areas of the scene. The experimental results show the performance of our proposal over the proposed evaluation dataset PDds. There is a global detection improvement in almost every category and original people detection approach, being this improvement more clear in those scenarios with medium or high background complexity, since those scenarios are more likely to generate false detections. The results also show how the use of motion in addition to our approach obtains the best final results.

In the following chapter, we explore a different post-processing subtask: we propose to combine different detection approaches in order to add robustness to the detection and, therefore, improve the detection results.

Chapter 8

Decision-level fusion of people detectors

8.1 Introduction[1]

As already mentioned, people detection is one of the most challenging problems in computer vision. People detection approaches from the state of the art obtain satisfactory results in low and medium complexity scenarios, but these results are considerably reduced in more complex and realistic scenarios (see chapter 4). In order to achieve a more reliable performance in complex scenarios, we have proposed a new people detection approach based on motion and their combination with appearance information (see chapter 5) and the integration of this appearance and motion information in a detection and tracking system that takes advantage of the tracking information (see chapter 6). In the previous chapter, we have also proposed a new people detection post-processing subtask that reduces one of the main problems of people detection and, therefore, improves the global detection results (see chapter 7). In this chapter, we propose to combine different detection approaches in order to add robustness to the detection.

The main contribution presented in this chapter is a comprehensive study of different people detection approaches from the state of the art and their combination at decision-level in order to take advantage of their independent strengths and, at the same time, reduce their drawbacks and limitations; therefore, improving the global detection performance in typical video surveillance environments.

In this chapter, we will firstly make a brief introduction to the related literature in section 8.2. Then, the proposed combination at decision-level of multiple people detectors is described in section 8.3. After that, section 8.4 describes the experimental results. Finally, section 8.5 summarizes the chapter with some conclusions.

[1]This chapter is based on the publication *"A. García-Martín, J. M. Martínez. Decision-level fusion of appearance-based people detectors. Submitted to Electronic Letters"*

8.2 Related work

The combination or fusion of multiple information sources (multisensor, multimodality, etc) has been already thoroughly studied in the literature. Any fusion technique attempts to combine the information from all available sources into a unified representation that provides better information for human or machine perception as compared to any of the input sources. Several models for data fusion have been proposed in the literature. However, one of the models most commonly used in image processing applications is the three-level fusion model that is based on the levels at which information is represented [Hall and Llinas, 2001]. This model classifies data fusion into three levels: data or pixel-level fusion, feature fusion and decision fusion. At the lowest level, the fused pixel is derived from a set of pixels from the multiple input sources. At the intermediate level, the features for each object are independently extracted in each information source; these features create a common feature space for object classification. Finally, at the highest level, decision-level fusion corresponds to combining decisions from several experts.

In the case of people detection, every people detector must build up (explicitly or implicitly) some form of dense confidence map [Breitenstein et al., 2010], which consists of the continuous detection confidence score for each location and scale. [Felzenszwalb et al., 2010] combines or fuses the confidence map of several independent body parts at pixel-level in order to obtain a final confidence map that is used to localize people in the scene. Every people detector must design and train (if training is required) a person model based on characteristic parameters (motion, dimensions, silhouette, etc). There are some approaches that combine or fuse more than one feature at feature-level in order to improve the detection results: [Viola et al., 2003; Dalal and Triggs, 2006] combine appearance and motion expanding previous features based on appearance to more than one frame, whilst [Gan and Cheng, 2011] uses the feature HOG-LBP (combination of the HOG [Dalal and Triggs, 2005] and LBP , Local Binary Patterns [Ojala and Pietikainen, 2002], features). Finally, every people detection must compare the previously defined or trained person model with the input image (or sequence) and make a final decision according to a similarity criteria. There are some approaches that combine or fuse multiple detectors at decision-level using multi body part detectors [Wu and Nevatia, 2005], multiple independent evidences [Fernández-Carbajales et al., 2008] or detectors (see chapter 5).

In this work, we combine or fuse up to six independently appearance based people detectors from the state of the art and their combination with our motion based people detector (see chapter 5) at decision-level. All detectors or experts are run in parallel and the final decision is obtained as a combination of local expert responses using fusion methods widely studied in the literature, but adapted to the particular case of people detection fusion at decision-level [Kuncheva, 2002]: average, product, minimum, maximum, median and majority vote.

8.3 People detectors fusion

This section enumerates the different people detection approaches used from the state of the art and the decision-level fusion proposed in order to improve the detection performance in typical video surveillance environments.

In this work, we propose the fusion of six independent appearance based people detectors from the state of the art and their combination with our motion based people detector: Edge (see chapter 4), Fusion [Fernández-Carbajales et al., 2008], HOG [Dalal and Triggs, 2005], ISM [Leibe et al., 2005], TUD [Andriluka et al., 2009], DTDP [Felzenszwalb et al., 2010] and IMM (see chapter 5).

In general, any people detector outcome l ($l = 1, ..., L$) always consists of a list of N detections in each frame t. Each detection n ($n = 1, ..., N$) is represented by its position (x, y) and dimensions (w, h) (i.e., bounding box or blob $B_n(x, y, w, h)$) and a People-detection Confidence PC_n ($0 \leq PC_n \leq 1$). In order to combine or fuse the different detectors, firstly it is necessary to find matches or correspondences between every people detection from one detector B_n^l with the detections from the other detectors $B_n^{q \neq l}$, the chosen matching criteria is the Multiple Hypotheses Simplification Criteria ($MHSC$ -see section 5.3.2.3-). The $MHSC$ allows us to compare hypotheses at different scales using the three evaluation criteria defined by [Leibe et al., 2005]: the relative distance, cover and overlap. The relative distance (dr) measures the distance between the bounding box centers in relation to their size. Cover and overlap measure how much of one bounding box hypothesis is covered by the other and vice versa. A matching is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

Every people detector l has generally a different outcome N^l in each frame t, the number of detections and the detections themselves are not always matched between approaches (there is no unequivocal relationship between detectors' outcomes), so we are not able to apply directly the traditional fusion techniques [Kuncheva, 2002]: average, product, minimum, maximum, median and majority vote. For this reason, we evaluate the four first mentioned fusion techniques, but taking also into account the minimum number of matches required in the fusion (variation of majority vote) in order to validate the fusion. Therefore, we perform the fusion and evaluate the four fusion techniques for each possible number of matches m ($m = 1, ..., L$) (assuming that one match corresponds actually to no matching, i.e., the detection is presented in only one detector). The final outcome is again a list of N^{out} detections, where each detection n ($n = 1, ..., N^{out}$) is represented by the matched averaged bounding box B_n^{out}, the final People-detection Confidence resulting to apply the corresponding fusion technique PC_n^{out} ($0 \leq PC_n^{out} \leq 1$) and the corresponding number of matches for each final detection m_n^{out}. Each final bounding box B_n^{out} is obtained as the average of the respective matched bounding boxes, whilst each final People-detection Confidence PC_n^{out} is obtained applying the corresponding fusion technique over the

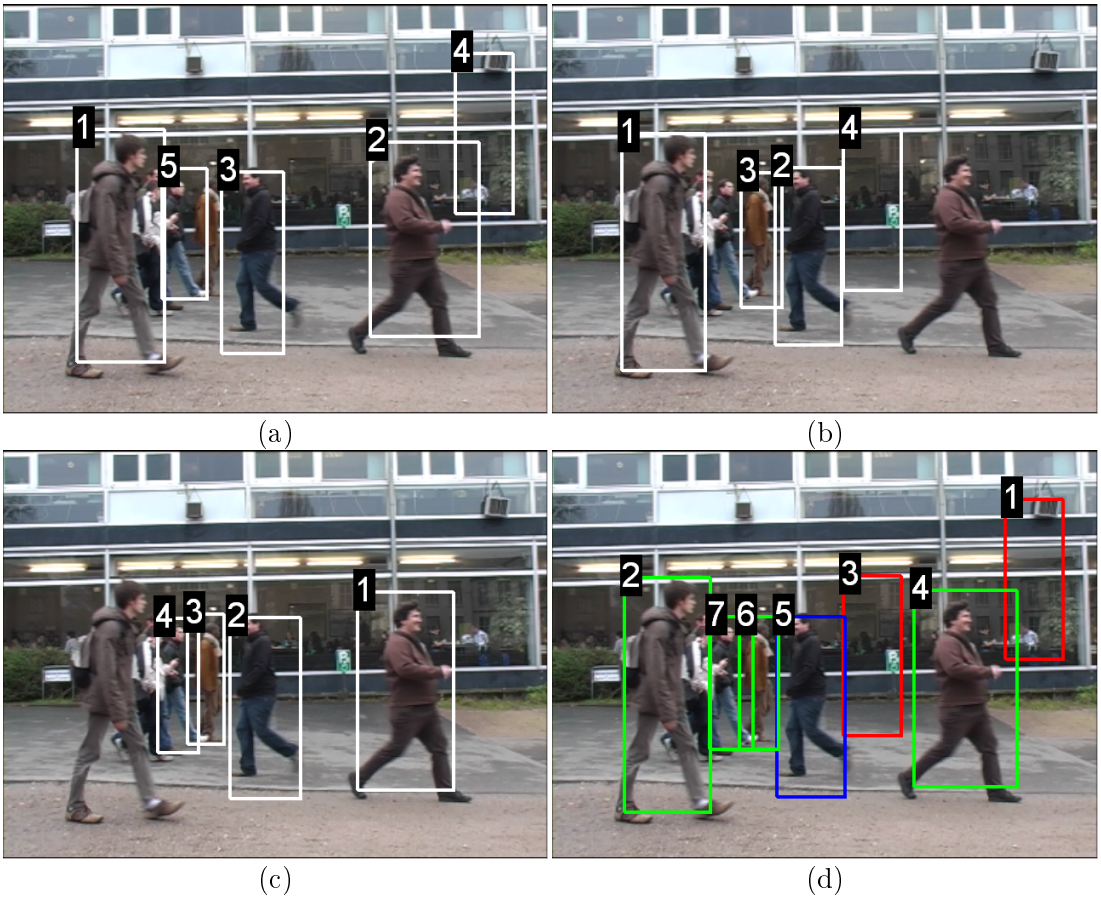


Fig. 8.1. Visual people detection fusion example: (a) people detector outcome l^1; (b) people detector outcome l^2; (c) people detector outcome l^3; and (d) final people detection fusion outcome l^{out} (see Algorithm 8.1). Blue color corresponds to $m_5^{out} = 3$, green color corresponds to $m_{2,4,6,7}^{out} = 2$ and red color corresponds to $m_{1,3}^{out} = 1$.

People-detection Confidence of the respective matched bounding boxes. Figure 8.1 and pseudo code algorithm 8.1 show a fusion example with three detectors.

8.4 Experimental results

This section describes the experimental setup used in the evaluation of the proposed people detection fusion at decision-level, the experimental results and the computational cost.

8.4.1 Experimental setup

In order to evaluate our people detection approach, we compare in this section the original performance and the fusion of seven independent people detectors from the state of the art: Edge (see chapter 4), Fusion [Fernández-Carbajales et al., 2008], HOG [Dalal and Triggs, 2005],

Algorithm 8.1 People detection fusion example.

- $L = 3$ $(l = 1, 2, 3)$.
- $l^{out} = fusion \begin{cases} l = 1, N^1 = 5 & \left\{B_1^1, PC_1^1\right\}, ..., \left\{B_5^1, PC_5^1\right\}. \\ l = 2, N^2 = 4 & \left\{B_1^2, PC_1^2\right\}, ..., \left\{B_4^2, PC_4^2\right\}. \\ l = 3, N^3 = 4 & \left\{B_1^3, PC_1^3\right\}, ..., \left\{B_4^3, PC_4^3\right\}. \end{cases}$
- $N^{out} = 7, l^{out} = \left\{B_1^{out} = B_4^1, PC_1^{out} = PC_4^1, m_1^{out} = 1\right\}, ...$
  $\left\{B_7^{out} = \frac{\left(B_5^1 + B_4^3\right)}{2}, PC_7^{out} = fusion^* \left(PC_5^1, PC_4^3\right), m_7^{out} = 2\right\}.$

$^*average, product, minimum, maximum$ *or* $median.$

ISM [Leibe et al., 2005], TUD [Andriluka et al., 2009], DTDP [Felzenszwalb et al., 2010] and IMM (see chapter 5). There is a brief description of the different people detection approaches used from the state of the art in appendix A.

As in the previous chapter 7, focused on the idea of evaluating the performance of the proposed approach in different typical video surveillance environments, it has been evaluated in both evaluation datasets (A and B) described in the performance evaluation methodology (see section 3.3). The dataset A allows us to evaluate the different approaches at every complexity level (C1,..,C5), whilst the dataset B allows us to evaluate more thoroughly the highest complexity category (C5). The different detection approaches experimental results have been obtained using the available code and binaries.

8.4.2 People detection results

8.4.2.1 Evaluation dataset A

Firstly, we evaluate and compare the six independently appearance based people detectors from the state of the art at every complexity level using the evaluation dataset A. The original people detection results have been already discussed in chapter 4.

According to the original people detection results (see section 4.4.2.1), we have defined two different people detection fusion configurations: the first one including the six detectors in the fusion and the other one without the detector with the worst detection results (Fusion). Additionally, we have evaluated every possible minimum number of matches m $(m = 1, ..., L)$ required in the fusion. Figure 8.2(a) shows the average results fusing the six detectors over the five experimental dataset complexity categories (C1-C5), whilst Figure 8.2(b) shows the same results, but fusing only five detectors (without Fusion detector). Firstly, in both cases it is clear the effect of the minimum number of matches required in the fusion. With low concurrence requirements $m = 1$ or high concurrence requirements $m = 6/5$ the final results are clearly worse. In the first case, it is because every detection is considered in the fusion, so every independent and isolated

$m = 3$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C1	0.99	+1.0	+26.9	+7.6	+4.2	+6.5	+3.1	+8.2
C2	0.96	+3.2	+18.5	+11.6	+5.5	+9.1	+4.3	+8.7
C3	0.86	+1.2	+43.3	+16.2	+7.5	+14.7	+6.2	+14.8
C4	0.90	+1.1	+30.4	+9.8	+7.1	+7.1	+4.7	+10.0
C5	0.77	+10.0	+60.4	+8.5	+8.5	+14.9	+4.1	+17.7
Total	0.90	+3.3	+35.9	+10.7	+6.6	+10.5	+4.5	+11.9

Table 8.1: People detection performance fusing the six detectors using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A. Percentage increase (%Δ) calculated with respect to original individual performance (see section 4.4.2.1).

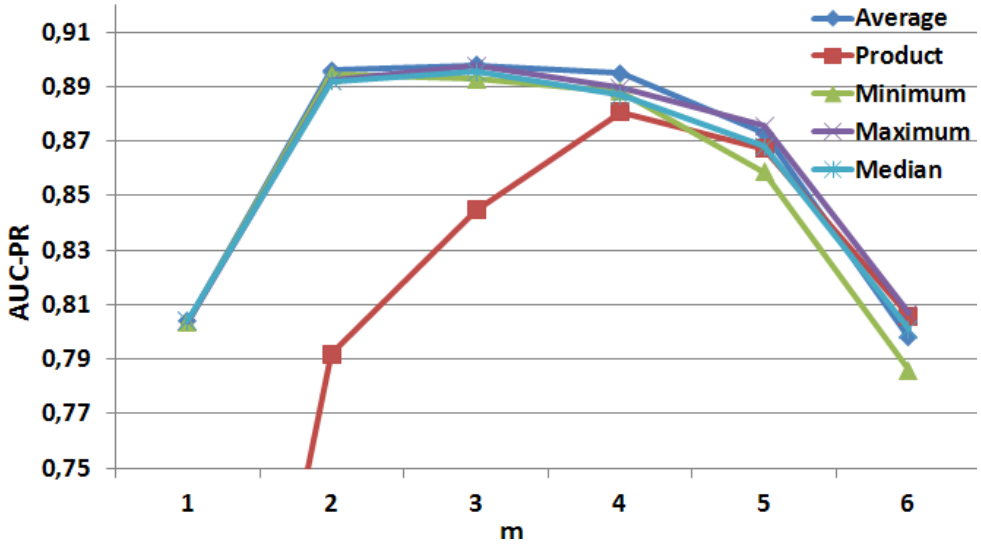
$m = 2$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C1	1.0	+2.0	+28.2	+8.7	+5.3	+7.5	+4.2	+9.3
C2	0.97	+4.3	+19.8	+12.8	+6.6	+10.2	+5.4	+9.9
C3	0.87	+2.4	+45.0	+17.6	+8.8	+16.0	+7.4	+16.2
C4	0.92	+3.4	+33.3	+12.2	+9.5	+9.5	+7.0	+12.5
C5	0.81	+15.7	+68.8	+14.1	+14.1	+20.9	+9.5	+23.8
Total	0.91	+5.6	+39.0	+13.1	+8.8	+12.8	+6.7	+14.3

Table 8.2: People detection performance fusing the five detectors (without Fusion detector [Fernández-Carbajales et al., 2008]) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A. Percentage increase (%Δ) calculated with respect to original individual performance (see section 4.4.2.1).

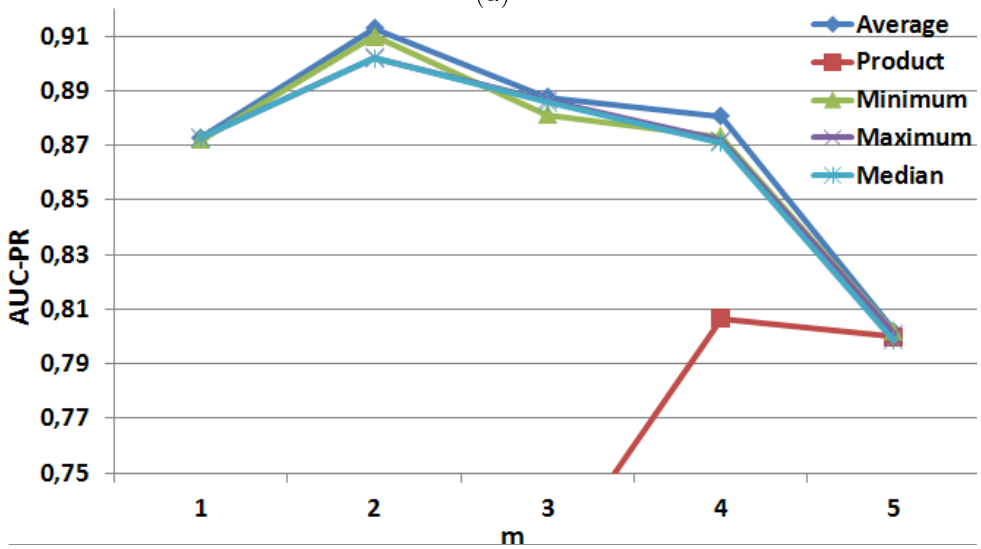
detection error is included in the final results. In the second case, there are missing detections due to the excessive detection concurrence requirements. The best results are obtained around $m = 3$ or $m = 2$ respectively. In relation to the chosen fusion technique, the product method gets clearly the worst fused results: the product method is optimal only if all the detectors are totally independent. Although all the detectors are independently build, there is some kind of dependence since all of them are based on people appearance. The rest of fusion methods get similar results, being slightly worse the minimum and slightly better the average for every possible minimum number of matches required in the fusion.

According to the average results fusing the six or five detectors and in order to visualize the detection results per each experimental dataset complexity category, we have selected the best number of minimum matches required for each configuration (six detectors fusion $m = 3$ and only five detectors $m = 2$) and we have selected only the best performance fusion method (average). All experimental results are available as additional material (see Appendix B).

Table 8.1 shows the people detection performance fusing the six detectors per each experi-



(a)



(b)

Fig. 8.2. Total average fusion performance in terms of area under the Precision-Recall curve (AUC-PR) of dataset A, for each fusion technique [Kuncheva, 2002] (average, product, minimum, maximum and median) and minimum number (m) of matches required in the fusion: (a) fusing the six detectors and (b) fusing the five detectors (without Fusion detector [Fernández-Carbajales et al., 2008]).

mental dataset complexity, whilst Table 8.2 shows the same results, but fusing only five detectors (without Fusion detector). Both results show clearly that the proposed people detection fusion improves considerably the original people detection results. The average improvements obtained for each experimental dataset complexity are between 8.2 and 17.7% fusing six detectors and between 9.3 and 23.8% fusing only five. Although there is not a great difference between the two configurations, the results show how the use of a clearly worse original detector reduces the fusion improvements. Finally, the average improvements obtained are clearly higher in more complex scenarios (C3-C5) than in the simplest ones (C1-C2). It is logical because the range of possible improvement is greater and it is more evident the advantage of combining detectors (allowing to reduce errors and increase the overall detection rate).

According to the individual people detector results, the improvements on those detectors with worse original performance are logically greater than the improvements on those detectors with better original performance. On the one hand, the Fusion approach gets the worst original performance results (see section 4.4.2.1) and the greatest improvement (average improvement between 35.9∼39.0%). On the other hand, the Edge detector gets the best original performance results (see section 4.4.2.1) and the lowest improvement (average improvement between 3.3∼5.6%).

8.4.2.2 Evaluation dataset B

In this section, we evaluate more thoroughly the highest complexity category (C5) of the six independently appearance based people detectors from the state of the art using the dataset B. The original people detection results have been already discussed in chapter 4.

According to the original people detection results (see section 4.4.2.2), we have defined three different people detection fusion configurations: the first one including the six detectors in the fusion, the second one without the detector with the worst detection results (Fusion) and the third one including only the three best detectors (HOG, ISM and DTDP).

As in the evaluation of dataset A, in order to visualize the detection results, we have selected the best number of minimum matches required for each configuration (six detectors fusion $m = 3$, only five detectors $m = 2$ and only the best three detectors $m = 2$) and we have selected only the best performance fusion method (average). All experimental results are available as additional material (see Appendix B).

Table 8.3 shows the people detection performance fusing the six detectors, Table 8.4 shows the same results, but fusing only five detectors (without Fusion detector), whilst Table 8.5 shows the same results, but fusing only the three best detectors (HOG, ISM and DTDP). In almost all the cases we improve the global detection results: we can see how the improvements obtained fusing only the best three detectors (total average improvement of 22.3%) are significantly better than the ones obtained fusing the six or five detectors (total average improvement of 17.2 and 18.9% respectively).

$m = 3$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.69	+16.9	+56.8	+4.6	+0.0	+23.2	+1.5	+17.2

Table 8.3: People detection performance fusing the six detectors using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (%Δ) calculated with respect to original individual performance (see section 4.4.2.2).

$m = 2$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.70	+18.6	+59.1	+6.1	+1.4	+25.0	+2.9	+18.9

Table 8.4: People detection performance fusing the five detectors (without Fusion detector [Fernández-Carbajales et al., 2008]) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (%Δ) calculated with respect to original individual performance (see section 4.4.2.2).

In relation to the individual people detector results, on the one hand, the Fusion approach gets the worst original performance results (see section 4.4.2.2) and the greatest improvement (average improvement between 56.8$\sim$63.6%). On the other hand, the ISM detector gets the best original performance results (see section 4.4.2.2) and the lowest improvement (average improvement between 0.0$\sim$4.3%).

8.4.2.3 Evaluation dataset B with motion

In this section, we evaluate again the dataset B with the six independently appearance based people detectors from the state of the art, but in this case including the people detector based on motion IMM and all the appearance and motion combinations (Edge+IMM, Fusion+IMM, HOG+IMM, ISM+IMM, TUD+IMM and DTDP+IMM). The original people detection results have been already discussed in previous chapter 7.

According to the original people detection results (see section 7.5.2.3) and following the same evaluation scheme as in the evaluation of dataset B (see previous section 8.4.2.2), we have defined the same three different people detection fusion configurations and the same evaluation parameters (average fusion method and minimum matches required for each configuration). All experimental results are available as additional material (see Appendix B).

$m = 2$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total
C5	0.72	+22.0	+63.6	+9.1	+4.3	+28.6	+5.9	+22.3

Table 8.5: People detection performance fusing the three detectors (HOG, ISM and DTDP) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (%Δ) calculated with respect to original individual performance (see section 4.4.2.2).

$m=3$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.68	+17.2	+47.8	+3.0	+6.3	+21.4	+1.5	+16.2	+13.3

Table 8.6: People detection performance fusing the six appearance based detectors using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

$m=2$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.70	+20.7	+52.2	+6.1	+9.4	+25.0	+4.5	+19.6	+16.7

Table 8.7: People detection performance fusing the five appearance based detectors (without Fusion detector [Fernández-Carbajales et al., 2008]) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

Tables 8.6, 8.7 and 8.8 show the people detection performance fusing six, five or three appearance based detectors respectively and the motion based detector performance. In almost all the cases we improve the global detection results: we can see how the improvements obtained fusing only the best three detectors (total average improvement of 21.3%) are significantly better than the ones obtained fusing the six or five detectors (total average improvement of 16.2 and 19.6% respectively) and are quite similar than the ones obtained with the dataset B with motion (see previous section 8.4.2.2).

According to the individual people detector results, on the one hand, the Fusion approach gets the worst original performance results (see section 7.5.2.3) and the greatest improvement (average improvement between 47.8∼54.3%). On the other hand, the DTDP detector gets the best original performance results (see section 7.5.2.3) and the lowest improvement (average improvement between 1.5∼6.0%).

Tables 8.9, 8.10 and 8.11 show the people detection performance fusing six, five or three appearance and motion based detectors combinations respectively. Again, in almost all the cases we improve the global detection results, we can see how the improvements obtained fusing only the best three detectors (total average improvement of 17.5%) are significantly better than the ones obtained fusing the six or five detectors (total average improvement of 12.7 and 15.9% respectively).

The Fusion+IMM approach gets the worst original performance results (see section 7.5.2.3) and the greatest improvement (average improvement between 42.9∼49.0%). On the other hand, the DTDP+IMM detector gets the best original performance results (see section 7.5.2.3) and the lowest improvement (average improvement between 0.0∼4.3%).

Finally, the results show how the use of motion in addition with the proposed fusion obtains the best final results (AUC-PR final between 70 or 73%).

$m = 2$	Edge	Fusion	HOG	ISM	TUD	DTDP	Total	IMM
C5	0.71	+22.4	+54.3	+7.6	+10.9	+26.8	+6.0	+21.3	+18.3

Table 8.8: People detection performance fusing the three appearance based detectors (HOG, ISM and DTDP) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

$m = 3$	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.70	+12.9	+42.9	+2.9	+4.5	+12.9	+0.0	+12.7

Table 8.9: People detection performance fusing the six appearance and motion based detectors combinations using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

$m = 2$	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.72	+16.1	+46.9	+5.9	+7.5	+16.1	+2.9	+15.9

Table 8.10: People detection performance fusing the five appearance and motion based detectors combinations (without Fusion+IMM detector) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

$m = 2$	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	Total
C5	0.73	+17.7	+49.0	+7.4	+9.0	+17.7	+4.3	+17.5

Table 8.11: People detection performance fusing the three appearance and motion based detectors combinations (HOG+IMM, ISM+IMM and DTDP+IMM) using average fusion, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion. Percentage increase (%Δ) calculated with respect to original individual performance (see section 7.5.2.3).

8.4.3 Computational cost

According to the computational cost, running all people detectors in parallel and being almost insignificant the computational cost of the matching and fusion between detectors in comparison with the computational cost of the detectors, the final fusion approach computational cost will be established by the detection approach with the higher computational cost.

8.5 Summary and conclusions

We have presented a comprehensive study of different people detection approaches from the state of the art and the combination or fusion of six independent appearance based people detectors at decision-level and their combination with our motion based people detector in order to improve the detection performance in typical video surveillance environments. In order to fuse the different detectors, we have presented a multi people detection combination criteria and the application of traditional fusion techniques: average, product, minimum, maximum and median. The experimental results show the performance of our proposed fusion with the mentioned fusion techniques. The product method shows clearly worse results, whilst the average method gets slightly better results than the other three methods. The experimental results also show the performance of our proposal over the proposed evaluation dataset PDds. There is a global detection improvement in every category and original people detection approach, being this improvement more clear in those scenarios with higher complexity, since those scenarios are more likely to generate false detections and missing detections. Finally, the results show how the use of motion in addition with the proposed fusion obtains the best final results.

Part III

Conclusions

Chapter 9

Achievements, conclusions and future work

9.1 Summary of achievements and main conclusions

This thesis has addressed people detection in video surveillance scenarios. The goal was to analyze the most representative approaches from the state of the art, identify their weaknesses and propose contributions to improve current people detection approaches. In particular, two areas have been explored related with people detection benchmarking (chapter 3) and people detection approaches (chapters 4, 5, 6, 7).

In the first part of this thesis, we have described the motivations and considerations applied to the generation of a corpus (dataset and associated ground-truth) and the definition of a performance evaluation methodology for the evaluation of people detection algorithms in video sequences (chapter 3). A more complete people detection corpus in surveillance scenarios than the ones available in the state of the art has been produced (Person Detection dataset or PDds). Both the wide range of considered critical factors and the development of an accurate ground-truth for the presented corpus, makes it especially suitable for tuning the algorithms, results evaluation and comparison. A people detection evaluation methodology has been defined with a particular interest in assessing the overall detection system performance instead of just the binary classifier performance (person/no person). Altogether, a complete framework for the evaluation of people detection algorithms under different complexity conditions has been provided.

In the second part of this thesis, we have proposed three different people detection algorithms. Firstly, we have proposed a people detection approach that combines both initial object hypotheses generation or extraction techniques, i.e., segmentation and exhaustive search, in order to achieve robustness and real time operation (chapter 4). A complete surveillance video system has been implemented to evaluate the proposed detection approach. Besides, in order

to provide a good performance evaluation of the proposed framework, it has been evaluated over the proposed evaluation dataset PDds. Experimental results over the proposed evaluation dataset A show that the proposed system performs considerably well at real time and even better than other non-real time approaches from the state of the art and that it is significantly more efficient and stable than others approaches from the state of the art. However, due to the background segmentation difficulty in complex scenarios, at high levels of complexity our proposal obtains similar results than the state of the art. Experimental results over the proposed evaluation dataset B points out that our approach does not work properly in more complex and realistic scenarios. Our approach presents a strong dependence with the segmentation stage, so all the segmentation problems are inherited (under and over segmentation). Our combination of segmentation and exhaustive search reduce these problems, but these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation.

Secondly, we proposed a people detection approach that combines an appearance people model from the state of the art and our motion people model (chapter 5). Using the ISM Framework and the MoSIFT interest points detector and descriptor, we present a new people detection algorithm based in the characteristic movements of people. Experiments have been conducted on challenging and realistic sequences extracted from the TRECVID dataset and part of our evaluation dataset PDds with the maximum complexity category. The results show that our motion-based detector produces results comparable to the ISM state of the art approach in complex and realistic scenarios and, therefore, the human motion provides useful information for people detection and independent from appearance information. The evaluation of the whole system shows how the combination of different information sources improves the final detection, obtaining a significant improvement in Recall and a slightly Precision reduction.

In the third place, this thesis has explored to take advantage of the appearance and motion combination over time with a collaborative people detection and tracking system (chapter 6). It integrates the people detection and tracking information into a single system and improves both tasks simultaneously. We have analyzed the different system configurations in order to evaluate the improvement introduced by the mutual information exchange. Experiments have been conducted on challenging sequences extracted from our evaluation dataset PDds created with TRECVID sequences (highly crowded scenes, severely cluttered background and people at different scales), highlighting the problems that these complex scenarios entail in the state of the art of people detection and tracking. The experiments on the proposed dataset show the utility of the collaborative system, specially in complex scenarios, getting better results than the state of the art for each task independently. The detection and tracking modules can be replaced by others without great difficulty thanks to the modular design of the system that allows a collaborative or independent performance, the generic format of the information to be exchanged (blobs and detection/tracking confidence) and the easily compatible information exchange mechanism (simple

and consistent process updates). The use of different modules will vary the overall performance of the system, but the combination of both sources of information, in principle, will be useful for improving the system (except in the ideal case of perfect detection and perfect tracking). With respect to people detection, firstly, we have used a people detector based on the combination of appearance and motion information. We have evaluated different appearance-motion combinations of people detectors from the state of the art and it is clear that human motion provides useful information for people detection and independent from appearance information. Secondly, a people detection prediction or update scheme using the tracking information about our collaborative system has been proposed and all the different people detector variations have been re-evaluated. The experimental results show that the use of tracking information stabilizes the people detection over time, so there is a significant improvement mainly in terms of Recall and F1Score. With respect to tracking, in a first place, an adaptive particle filter tracker based on color distributions with different people detection initializations has been evaluated. All trackers follow a similar pattern, but it is shown clearly that the initialization has a great influence on the global tracker performance. Secondly, all the tracker variations have been re-evaluated adding the people detection information about our collaborative system. The experimental results show that the use of people detection information corrects the position, dimension and color distribution of the trackers over time, so there is a significant improvement mainly in terms of Precision and F1Score.

In addition, also in the second part of this thesis, we have proposed two different people detection post-processing subtasks. Firstly, we have proposed a people-background segmentation approach that aims to ensure that there are no people or body parts assigned to the background class at the cost of potentially increasing the number of background pixels classified as people and, then, we have proposed a new people detection post-processing subtask based on this people-background segmentation (chapter 7). The experimental results show the performance of our proposal over the proposed evaluation dataset PDds. There is a global detection improvement in almost every category and original people detection approach, being this improvement more clear in those scenarios with medium or high background complexity, since those scenarios are more likely to generate false detections. Secondly, we have also proposed the combination or fusion of six independent appearance based people detectors at decision-level and their combination with our motion based people detector in order to improve the detection performance in typical video surveillance environments (chapter 8). In order to fuse the different detectors, we have presented a multi people detection combination criteria and the application of traditional fusion techniques: average, product, minimum, maximum and median. The experimental results show the performance of our proposed fusion with the mentioned fusion techniques. The product method shows clearly worse results, whilst the average method gets slightly better results than the other three methods. The experimental results also show the performance of our proposal

over the proposed evaluation dataset PDds. There is a global detection improvement in every category and original people detection approach, being this improvement more clear in those scenarios with higher complexity, since those scenarios are more likely to generate false detections and missing detections. Finally, in both cases, the results also show how the use of motion in addition to the use of both proposed post-processing subtasks obtains the best final results.

9.2 Comparative analysis of proposed people detection approaches

As already commented, the main objective of this thesis was to explore the state of the art in people detection in surveillance scenarios, analyze the most representative approaches, identify their weaknesses and propose contributions to improve current people detection state of the art. For this reason, different people detection approaches have been proposed in order to solve several limitations from the state of the art. In this section, we will compare and summarize the main people detection issues covered for each proposed people detection approach: people detection algorithms and post-processing subtasks. Table 9.1 summarizes and compare the main advantages and limitations of each approach.

Our first objective was a robust and real time people detector (see chapter 4). So we proposed to combine the two object detection approaches (segmentation and exhaustive search) in order to get a robust detection but also performing in real time. The proposed approach performs considerably well at real time in low and medium complexity scenarios. However, our approach still presents a strong dependence with the segmentation stage, so it does not work properly in more complex and realistic scenarios.

Our second objective was a people detector working in more complex and realistic scenarios (chapter 5). So, we proposed a person model based on motion information and the combination of appearance a motion models. In this case, we make use of approaches based on exhaustive search, since they are more robust in complex scenarios. The proposed approach does not perform in real time, but it gets better results in more complex and realistic scenarios.

Our next objective was the correction of people detection "unstable" behavior over time (see chapter 6). So, we proposed a collaborative scheme to improve simultaneously people detection and tracking. Again, the proposed approach does not perform in real time, but it works properly in more complex and realistic scenarios and is able to stabilize the detection over time.

Finally, both proposed additional post-processing subtask were focused on reducing the critical people detection compromise (false positive vs. missing detections) and, therefore, improve the global people detection results. In the first case, the main objective was the reduction of false detection using our novel people-background segmentation (see chapter 7). In the second case, the main objective was the combination of independent people detectors strengths and the reduction of their limitations. However, in both cases, the main disadvantage is the additional

Approach	Advantages	Limitations
Chapter 4	Segmentation+Exhaustive search Real time Low and medium complexity scenarios	No complex scenarios Detection unstable over time False vs. missing detections balance
Chapter 5	Exhaustive search Appearance+Motion Complex scenarios	No real time Detection unstable over time False vs. missing detections balance
Chapter 6	Exhaustive search Appearance+Motion+Tracking Complex scenarios	No real time False vs. missing detections balance
Chapter 7	Reduces false detections Reduces false vs. missing detections balance	Additional computational cost
Chapter 8	Combines strengths and reduce drawbacks Reduces false vs. missing detections balance	Additional computational cost

Table 9.1: Comparative analysis of proposed people detection approaches.

computational cost of each post-processing subtask.

9.3 Future work

Based on the results and discussions of this thesis, we plan the following future research lines:

- Expand the evaluation dataset PDds. In chapter 3, the proposed experimental dataset PDds includes a great variability of scenarios with different background complexities and it also includes a great variability of people appearance and multiple interactions with objects and/or persons. However, we propose to extend the contents of the dataset and make use of every sequence recorded in a chroma studio and composed with every different background [Tiburzi et al., 2008], in order to be able to analyze independently the background and foreground factors.

- Improve or refine background subtraction. The people detector approach presented in chapter 4 combines segmentation and exhaustive search. As noted in the experimental results, our combination of segmentation and exhaustive search reduces the segmentation problems (under and over segmentation), but these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation. So, we propose the study of techniques for multimodal background modeling, noise removal, shadows detection, etc, in order to refine the background subtraction in complex scenarios.

- Appearance and motion fusion. In chapter 5, we proposed a people detection approach that combines two independent detectors: an appearance people model (ISM) from the

state of the art and our proposed motion people model (IMM). We propose the study of different fusion or combination techniques between the appearance and motion detectors to improve the Recall without compromising the Precision, or even the creation of a single integrated Implicit Shape-Motion Model (ISMM), using the full MoSIFT description.

- Expand tracking evaluation. The collaborative people detection and tracking system presented in chapter 6 has been tested with a particular tracker module. However, the tracking module can be replaced by others without great difficulty thanks to the modular design of the system that allows a collaborative or independent performance, the generic format of the information to be exchanged and the easily compatible information exchange mechanism. So, we propose the evaluation of the collaborative system with other trackers or even as in the case of the people detection, to combine efficiently multiple independent trackers.

- Forward and backward collaborative schemes. In chapter 6, we proposed a forward collaborative people detection and tracking system. We propose not to only deal with this kind of forward collaborative schemes, but also investigate forward and backward collaborative schemes, i.e., feedback systems.

- People-background segmentation. We propose to improve the people-background segmentation presented in chapter 7 incorporating temporal information in the model and explore the possibility of detecting automatically the range of scales presented in each part of the scene and the binarization threshold. In addition, we propose to extend the method to other people detector approaches and object classes.

- Segmentation confidence. In order to refine or improve the proposed segmentation confidence in chapter 7, we propose the combination of the people-background segmentation with another more traditional segmentation strategy: color based, motion based, etc. After showing that this post-processing allows improving detection results, we propose to study the use of the people-background segmentation as a preprocessing state in order to maintain or reduce computation cost. Finally, we also propose to explore other combinations of detection and segmentation confidences.

- Decision-level fusion of people detectors. In chapter 8, we proposed the combination or fusion of six independent appearance based people detectors and one motion based people detector. We propose to explore other more complex fusion possibilities, not only fixed fusion rules, but also trainable fusion rules or adaptive weights based on online quality estimation; and not only parallel fusion schemes, but also cascade, hierarchical or hybrid. Finally, it is clear that "independently build" detectors exhibit positive correlation and this is attributed to the fact that difficult parts of the decision space are difficult for all detectors. So we also propose to explore other fusion techniques robust to decision correlations.

Part IV

Appendixes

Appendix A

People detectors

A.1 Introduction

In this appendix, we will make a brief introduction of the different people detection approaches used from the state of the art.

A.2 People detectors

This section enumerates and describes briefly the different people detection approaches used from the state of the art: Edge (see chapter 4), Fusion [Fernández-Carbajales et al., 2008], HOG [Dalal and Triggs, 2005], ISM [Leibe et al., 2005], TUD [Andriluka et al., 2009], DTDP [Felzenszwalb et al., 2010] and IMM (see chapter 5).

The Edge detector (see chapter 4) combines segmentation and exhaustive search in order to achieve robustness and real time operation. It is a real time adaptation of the people detection approach [Wu and Nevatia, 2005]. An individual human is modeled as an assembly of natural body parts. The main idea consists of identifying characteristic edges of each body part and generating four edge models of body parts (body, head, torso and legs). The initial objects candidates to be person are extracted using background subtraction and then those selected candidates are scanned with four independent edge feature detectors previously trained.

The Fusion detector [Fernández-Carbajales et al., 2008] is a real time detection approach based on segmentation and a holistic person model. The initial objects candidates to be person are extracted using background subtraction and the holistic person model is the combination or fusion at decision level of three simple person models: ellipse fitting [Xu and Fujimura, 2003], ghost [Haritaoglu et al., 1998] and aspect ratio.

The HOG detector [Dalal and Triggs, 2005] is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model, evaluating different detection windows with a classifier at multiple scales and

locations. The chosen person model is based on appearance information using the Histogram of Oriented Gradients.

The ISM detector [Leibe et al., 2005] is a generative model for object detection and has been applied to a variety of object categories including cars, motorbikes, animals and pedestrians. The ISM people detector is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching. The chosen person model is based on appearance information using the SIFT features.

The TUD people detector [Andriluka et al., 2009] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original ISM detector [Leibe et al., 2005] using pictorial structures. The appearance of body parts is modeled using densely sampled shape context descriptors and discriminatively trained Adaboost classifiers. As a result, it presents a strong discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts.

The DTDP detector [Felzenszwalb et al., 2010] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original HOG detector [Dalal and Triggs, 2005]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable body part is modeled as the original HOG detector [Dalal and Triggs, 2005].

Finally, the IMM detector (see chapter 5) is based on feature-based exhaustive. The chosen person model is based in the characteristic movements of people using the ISM Framework and the MoSIFT interest points detector and descriptor. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching.

Appendix B

People-background segmentation experimental results

B.1 Introduction[1]

This appendix describes the experimental results of the proposed people-background segmentation approach (see chapter 7). In this appendix, we will firstly describe the experimental setup in section B.2. Then, the experimental results are described in section B.3. Finally, section B.4 describes the computational cost.

B.2 Experimental setup

In order to evaluate our unequal-error-cost people-background segmentation approach (see section 7.3), we compare in this section the performance of the original algorithm [Felzenszwalb et al., 2010], the independent and dependent body parts approaches (IBP and DBP, respectively), their extended versions (IEBP and DEBP, respectively) and the proposed method DEBP-P[2].

We use a set of publicly available sequences with different complexities, including occlusions, scale variations, different point of views and moving cameras: tree outdoors sequences (TUD-Campus and TUD-Crossing from [Andriluka et al., 2008] and PETS2009[3]), three indoor sequences (TRECVID2008[4], PETS2006[5] and AVSS2007[6]) and three sequences with moving cam-

[1] This appendix is based on the publication *"A. García-Martín, A. Cavallaro, J. M. Martínez. People-background segmentation with unequal error cost. In Proc. of the IEEE International Conference on Image Processing, 2012"*

[2] Video results, ground-truth and additional data can be found at http://www-vpu.eps.uam.es/publications/PeopleBackgroundSegmentation

[3] `http://www.cvg.rdg.ac.uk/PETS2009/`

[4] `http://www.itl.nist.gov/iad/mig//tests/trecvid/2008/`

[5] `http://www.cvg.rdg.ac.uk/PETS2006/`

[6] `http://www.avss2007.org`

Sequence	GTF	ANP	PPP	Resolution
TUD-Campus	7/71	6.1	14.13	640x480
TUD-Crossing	21/201	6.2	9.55	640x480
TRECVID	6/103	9.1	9.38	720x576
PETS2006	6/1010	2.3	2.59	720x576
PETSS2009	6/443	6.5	2.15	768x576
AVSS	6/907	2	3.68	720x576

Table B.1: Description of the experimental dataset (Key. GTF: number of ground-truth frames per sequence; ANP: average number of people per ground-truth frame; PPP: percentage of pixels belonging to a person in the ground-truth).

eras from [Ess et al., 2008].

In order to quantify the error, we manually generated a segmentation ground-truth for selected frames of the first six sequences (see Table B.1). Note that the image border (whose width is half the size of a person on both sides of the image, i.e., 20 or 40 pixels according to the model scale in [Felzenszwalb et al., 2010]) is not considered in the quantitative evaluation. The visual results of these annotated first six sequences have been generated with the maximum binarization threshold for which there are no pixels of people misclassified as background, whilst the visual results of the three non-static camera sequences have been generated with the empirical binarization threshold of 0.8.

In order to evaluate our people-background segmentation approach, we are interested in sub-unit performance evaluation (pixel in this case) and classification performance instead of overall system performance. As already commented in section 3.3.2, the sub-unit performance is usually measured in terms of Receiver Operating Characteristics (ROC) and gives us information about the classification stage, while the Precision-Recall (PR) provides overall system performance information. Table B.2 shows the results in terms of AUC-ROC (area under the ROC curve) with different false positives penalty factors: 1, 2, 4 and 10. A penalty factor of 1 corresponds to traditional segmentation approaches, whilst higher factors give higher penalties to segmentations with pixels that correspond to a person and are incorrectly classified as background, i.e., a penalty factor of 2 corresponds to a twice penalty and so on.

B.3 People-background segmentation results

The results show that *dependent-part* approaches (DBP and DEBP) outperform *independent-part* approaches (IBP and IBP) due to the greater robustness provided by the combined body parts detections. The extended versions (IEBP and DEBP) are significantly better than their non-extended counterparts (IBP and DBP) due to the reduction of the number of false positives

	TUD-Campus				TUD-Crossing				TRECVID				PETS2006				PETS2009				AVSS			
False positive penalty factor	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10	1	2	4	10
Original[1]	.83	.75	.65	.51	.87	.81	.73	.63	.83	.74	.65	.50	.77	.68	.59	.46	.88	.82	.75	.64	.89	.83	.75	.63
IBP	.81	.74	.66	.56	.78	.68	.58	.44	.65	.53	.41	.27	.68	.56	.45	.31	.70	.58	.44	.28	.69	.57	.46	.32
IEBP	.84	.79	.72	.63	.84	.76	.68	.55	.74	.63	.52	.37	.74	.63	.52	.37	.80	.70	.58	.42	.77	.67	.56	.41
DBP	.93	.90	.85	.79	.93	.89	.84	.76	.85	.77	.68	.54	.85	.78	.70	.58	.86	.77	.66	.50	.90	.85	.78	.67
DEBP	**.95**	**.93**	**.90**	**.85**	**.95**	**.93**	**.90**	**.85**	**.92**	**.88**	.83	.74	.93	.89	.85	.77	**.98**	.96	.93	.87	.95	.93	.90	.83
DEBP-P	.93	.91	.88	.84	.94	.92	**.90**	**.85**	**.92**	**.88**	**.84**	**.77**	**.94**	**.91**	**.87**	**.80**	**.98**	**.98**	**.96**	**.95**	**.96**	**.95**	**.93**	**.90**

Table B.2: Area under the ROC curve (AUC-ROC) with different false positive penalty factors. [1]Original algorithm [Felzenszwalb et al., 2010].

(pixels that belong to a person incorrectly classified as background) without a substantial increase of false negatives (pixels that belong to the background incorrectly classified as people). Despite the fact that IBP and IEBP were initially designed to reduce false positives, the lack of dependency among parts generates many false negatives leading to worse performance compared to the corresponding original algorithm. Whilst the other approaches decrease drastically their performance with the increase of the penalty factor, the combination of dependent and extended body part approach DEBP has the lowest decrease and the best system performance (0.98∼0.74). Its post-processed version, the proposed approach DEBP-P, practically maintains the same performance and improves slightly the results for higher penalty factors (0.98∼0.77).

Figure B.1 and Figure B.2 show examples of static and non-static camera scenarios, respectively. Figure B.1 shows the performance of the original algorithm in terms of detection: we can see examples of missing detections or false detections (people only partially detected) in each scenario. The best results (0.98∼0.95) are obtained in the sequence PETS2009, due to the person model [Felzenszwalb et al., 2010]. Although the person model supports different body parts configurations (deformable part model), it favors people with arms and legs close to the body. In the case of the PETS2009 sequence, people are better suited to the model due to the far field view. However, in the other scenarios, the person model must be adapted to larger pose variations (higher body part deformation costs), getting worse results. The other factors that have influenced the results are the presence of shadows and reflections in TRECVID, PETS2006 and AVSS that makes the detection more difficult; and the greater scales variation in TRECVID and PETS2006 that makes the confidence map more complex and introduces more false body part detections that worsen the results. A separate analysis for each scale, as opposed to the current approach of combining first all the scales and then performing segmentation, could improve the results.

B.4 Computational cost

This section describes the computational cost of the proposed people-background segmentation approach. As already commented, the proposed people-background segmentation method is based on [Felzenszwalb et al., 2010] for detecting body parts and extends this representation by appropriately grouping them. Then, we fuse detection confidence maps according to regions that are expected to be covered by the body parts. The corresponding background segmentation mask is finally generated after binarization and post-processing (see chapter 7). For this reason and following the computational cost evaluation of the original approach [Felzenszwalb et al., 2010], we evaluate the computational cost in terms of seconds per frame. We compare the original computational cost with our different people-background segmentation approaches: IBP, IEBP, DBP, DEBP and DEBP-P (see section 7.3). The system has been implemented in Matlab using the original available code[7]. The tests have been performed on an Intel Core 2 Duo with a CPU frequency of 2.93 GHz and 3.21GB RAM.

Table B.3 shows the average computational cost of the original approach and our people-background segmentation approaches per each evaluation sequence and Table B.4 shows the computational cost increase of each approach with respect to original approach. Firstly, the results show that the IBP approach presents always and the IEBP in some cases lower computational cost than the original approach. It is due to that the highest computational costs in these cases correspond to the full body part and unlike all the other approaches, the IBP and IEBP do not use the full body part. Secondly, in general the results show how the extended versions (IEBP and DEBP) have logically a higher computational cost than the non-extended versions (IBP and DBP). And also the extended and post-processed version (DEBP-P) has a higher computational cost than the non post-processed version (DEBP). Finally, despite the fact that the final post-processed (DEBP-P) version has the higher computational cost (average computational cost increase of 1 second per frame), the results show how this computational cost increase is smaller in those sequences that include people with smaller scales (TRECVID, PETS2006 and PETS2009). It is due to that the highest computational costs in these cases correspond to the body parts confidence maps computation at multiple scales and not to the extension and post-processing tasks.

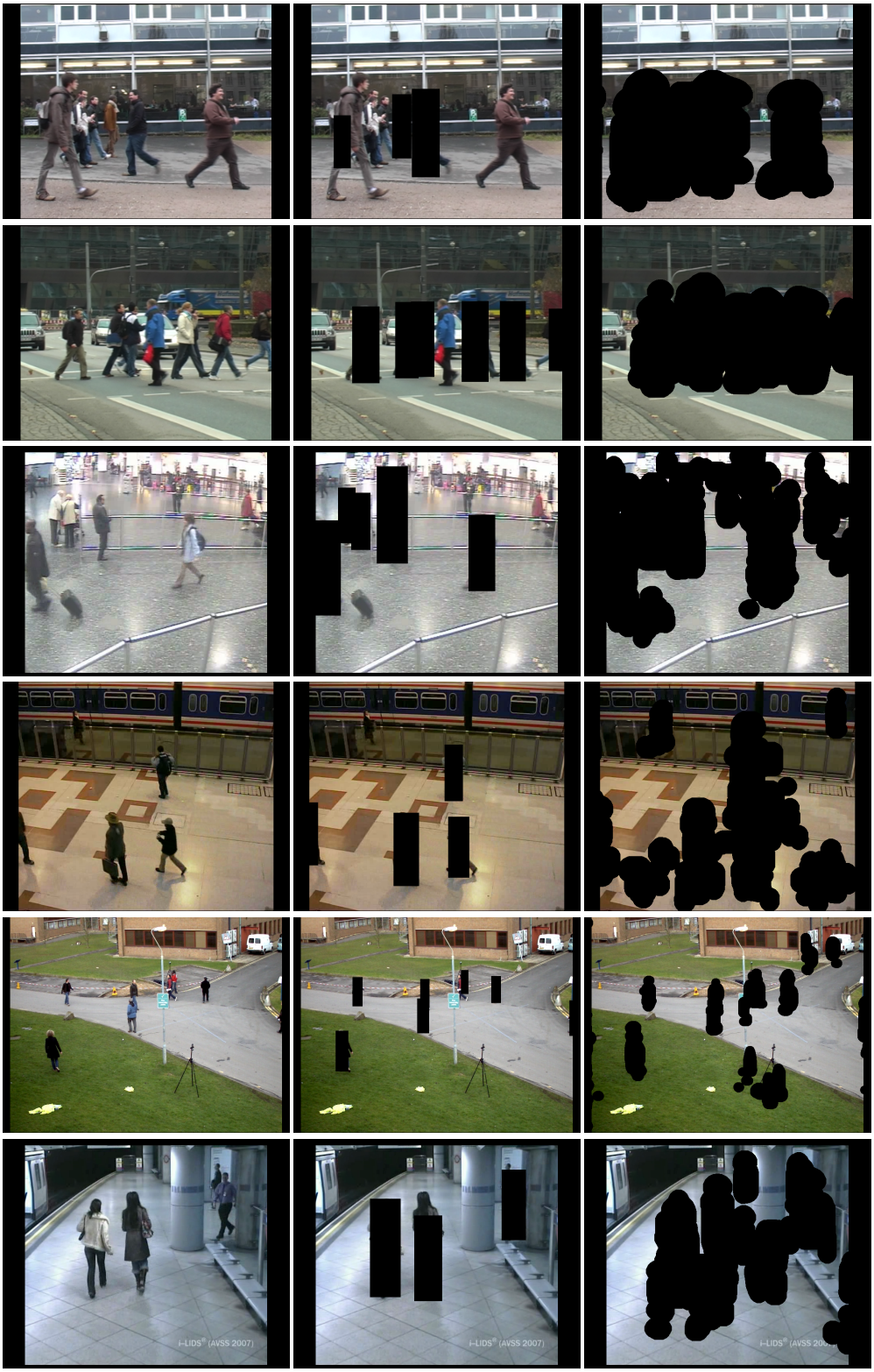
[7] `http://www.cs.brown.edu/~pff/latent/`

	TUD-Campus	TUD-Crossing	TRECVID	PETS2006	PETS2009	AVSS	Total
Original[1]	2.3	2.2	8.5	8.5	8.8	2.8	5.5
IBP	2.1	2.1	6.3	6.4	6.3	2.7	4.3
IEBP	2.4	2.4	6.8	6.8	6.7	3.0	6.7
DBP	3.0	3.0	9.0	9.0	9.0	3.9	6.2
DEBP	3.3	3.3	9.0	8.9	9.0	4.2	6.3
DEBP-P	3.5	3.5	9.3	9.2	9.2	4.5	6.5

Table B.3: Computational cost in seconds per frame.[1]Original algorithm [Felzenszwalb et al., 2010].

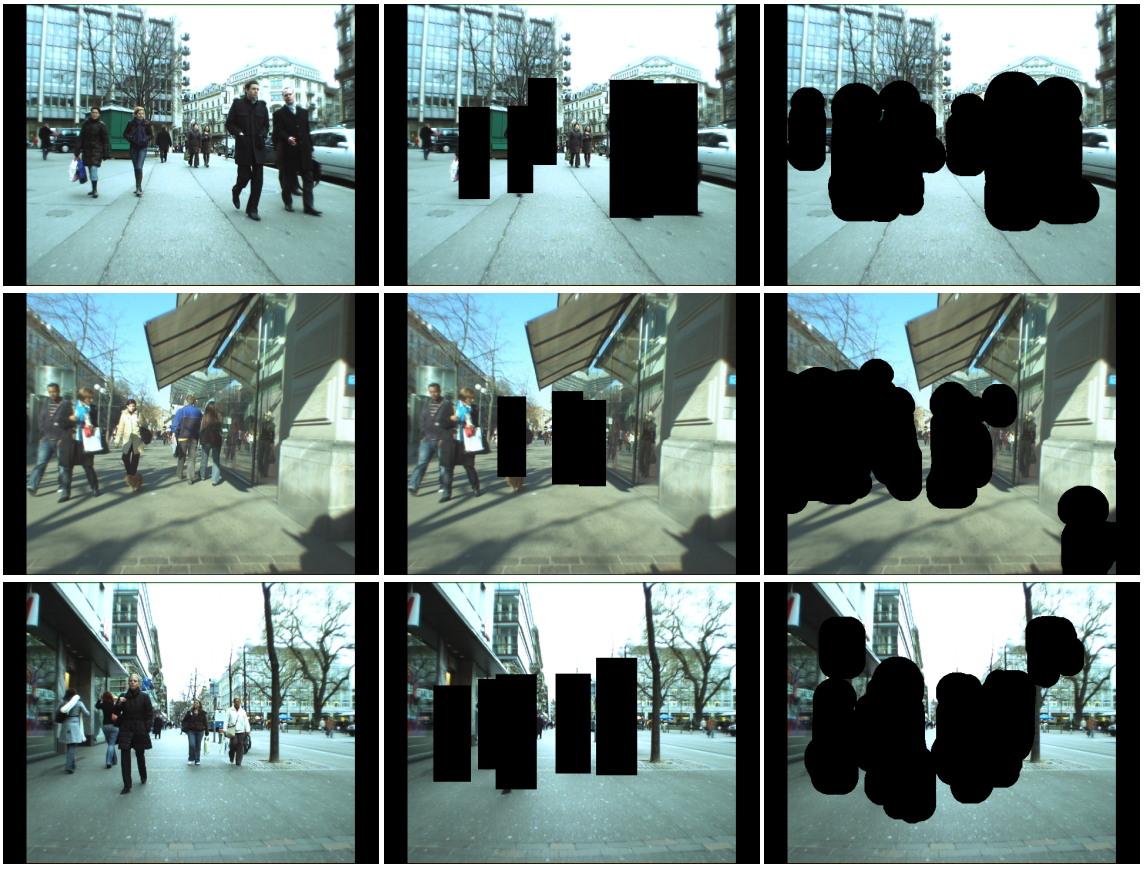
	TUD-Campus	TUD-Crossing	TRECVID	PETS2006	PETS2009	AVSS	Total
Original	-	-	-	-	-	-	
IBP	-0.2	-0.1	-2.2	-2.1	-2.5	-0.1	-1.2
IEBP	0.1	0.2	-1.7	-1.7	-2.1	0.2	-0.8
DBP	0.7	0.8	0.5	0.5	-0.2	1.1	+0.6
DEBP	1	1.1	0.5	0.4	0.2	1.4	+0.8
DEBP-P	1.2	1.3	0.8	0.7	0.4	1.7	+1.0

Table B.4: Computational cost increase (Δ) calculated with respect to original performance.



(a) (b) (c)

Fig. B.1. People-background segmentation sample results: (a) original frame; (b) person detector result; and (c) DEBP-P result.



(a) (b) (c)

Fig. B.2. People-background segmentation sample results for moving cameras: (a) original frame; (b) person detector result; and (c) DEBP-P background mask.

Appendix C

Decision-level fusion of people detectors additional experimental results

C.1 Introduction[1]

In the chapter 8, we presented the combination or fusion of six independent appearance based people detectors at decision-level and their combination with our motion based people detector in order to improve the detection performance in typical video surveillance environments. In order to fuse the different detectors, we presented a multi people detection combination criteria and the application of traditional fusion techniques: average, product, minimum, maximum and median. According to the average results of fusing the detectors and in order to visualize the detection results per each experimental dataset complexity category, in the chapter 8, we only selected the best number of minimum matches required for each configuration and we only selected only the best performance fusion method (average).

In this appendix, we will describe all the experimental results, including every number of matches required for each configuration and every fusion method.

C.2 Experimental results

This section describes all experimental results related with the proposed combination or fusion of six independent appearance based people detectors at decision-level and their combination with our motion based people detector (see chapter 8), including every number of matches required for each configuration and every fusion method.

[1]This appendix is based on the publication *"A. García-Martín, J. M. Martínez. Decision-level fusion of appearance-based people detectors. Submitted to IEEE International Conference on Advanced Video and Signal based Surveillance 2013"*

C.2.1 Evaluation dataset A

Table C.1 shows the people detection performance fusing the six appearance based detectors for each complexity category of evaluation dataset A, whilst Table C.2 shows the total average people detection performance fusing the six appearance based detectors of evaluation dataset A.

Table C.3 shows the people detection performance fusing the five appearance based detectors (without Fusion detector) for each complexity category of evaluation dataset A, whilst Table C.4 shows the total average people detection performance fusing the five appearance based detectors (without Fusion detector) of evaluation dataset A.

C.2.2 Evaluation dataset B

Table C.5 shows the people detection performance of dataset B fusing the six appearance based detectors, Table C.6 shows the same results but fusing only five appearance based detectors (without Fusion detector), whilst Table C.7 shows the same results but fusing only the three best appearance based detectors (HOG, ISM and DTDP).

C.2.3 Evaluation dataset B with motion

Table C.8 shows the people detection performance of dataset B with motion fusing the six appearance based detectors, Table C.9 shows the same results but fusing only five appearance based detectors (without Fusion detector), whilst Table C.10 shows the same results but fusing only the three best appearance based detectors (HOG, ISM and DTDP).

Table C.11 shows the people detection performance of dataset B with motion fusing the six appearance and motion based detectors combinations, Table C.12 shows the same results but fusing only five appearance and motion based detectors combinations (without Fusion+IMM detector), whilst Table C.13 shows the same results but fusing only the three best appearance and motion based detectors combinations (HOG+IMM, ISM+IMM and DTDP+IMM).

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
C1 Average	0.9099	0.9912	0.9948	0.9939	0.9810	0.9304
C1 Product	0.5776	0.9307	0.9767	0.9935	0.9816	0.9316
C1 Minimum	0.9099	0.9906	0.9916	0.9920	0.9778	0.9302
C1 Maximum	0.9099	0.9913	0.9933	0.9961	0.9874	0.9330
C1 Median	0.9099	0.9912	0.9948	0.9939	0.9810	0.9296
C2 Average	0.8820	0.9519	0.9643	0.9546	0.9226	0.8331
C2 Product	0.6444	0.8979	0.9440	0.9578	0.9398	0.8548
C2 Minimum	0.8817	0.9512	0.9633	0.9509	0.9171	0.8272
C2 Maximum	0.8820	0.9522	0.9644	0.9558	0.9273	0.8378
C2 Median	0.8819	0.9519	0.9625	0.9551	0.9229	0.8331
C3 Average	0.7630	0.8794	0.8639	0.8525	0.8313	0.7117
C3 Product	0.5400	0.7107	0.7769	0.8360	0.8375	0.7424
C3 Minimum	0.7629	0.8790	0.8660	0.8471	0.8150	0.7077
C3 Maximum	0.7629	0.8585	0.8499	0.8076	0.8052	0.7264
C3 Median	0.7630	0.8585	0.8500	0.8091	0.8065	0.7259
C4 Average	0.8455	0.8997	0.8916	0.9142	0.9021	0.8643
C4 Product	0.5298	0.8207	0.8530	0.9071	0.8766	0.8506
C4 Minimum	0.8461	0.9000	0.8885	0.9111	0.8807	0.8440
C4 Maximum	0.8450	0.8992	0.8919	0.9175	0.9117	0.8756
C4 Median	0.8457	0.9001	0.8921	0.9148	0.9002	0.8633
C5 Average	0.6195	0.7570	0.7710	0.7587	0.7272	0.6508
C5 Product	0.3803	0.5991	0.6731	0.7084	0.7004	0.6497
C5 Minimum	0.6171	0.7502	0.7524	0.7392	0.7020	0.6208
C5 Maximum	0.6201	0.7601	0.7800	0.7703	0.7457	0.6620
C5 Median	0.6196	0.7575	0.7722	0.7614	0.7296	0.6527

Table C.1: People detection performance fusing the six detectors, in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
Average	0.8039	0.8958	0.8971	0.8947	0.8728	0.7980
Product	0.5344	0.7918	0.8447	0.8805	0.8671	0.8058
Minimum	0.8035	0.8942	0.8923	0.8880	0.8585	0.7859
Maximum	0.8039	0.8922	0.8959	0.8894	0.8754	0.8069
Median	0.8040	0.8918	0.8943	0.8868	0.8680	0.8009

Table C.2: Total average fusion performance fusing the six detectors, in terms of area under the Precision-Recall curve (AUC-PR) of dataset A.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
C1 Average	0.9892	0.9984	0.9890	0.9841	0.9362
C1 Product	0.4333	0.7693	0.9119	0.9593	0.9433
C1 Minimum	0.9892	0.9984	0.9889	0.9855	0.9439
C1 Maximum	0.9892	0.9984	0.9885	0.9847	0.9367
C1 Median	0.9892	0.9984	0.9890	0.9841	0.9354
C2 Average	0.9142	0.9742	0.9609	0.9373	0.8354
C2 Product	0.4133	0.7538	0.8431	0.9123	0.8680
C2 Minimum	0.9138	0.9732	0.9585	0.9373	0.8470
C2 Maximum	0.9142	0.9745	0.9609	0.9368	0.8361
C2 Median	0.9142	0.9744	0.9611	0.9377	0.8350
C3 Average	0.8419	0.8679	0.8054	0.8010	0.6981
C3 Product	0.5897	0.4351	0.5146	0.6676	0.6997
C3 Minimum	0.8417	0.8669	0.8090	0.8003	0.7190
C3 Maximum	0.8417	0.8133	0.7973	0.7487	0.6746
C3 Median	0.8418	0.8135	0.7953	0.7495	0.6791
C4 Average	0.9060	0.9171	0.9030	0.9215	0.8835
C4 Product	0.4640	0.5977	0.7539	0.8663	0.8640
C4 Minimum	0.9056	0.9156	0.8952	0.9127	0.8654
C4 Maximum	0.9059	0.9161	0.9011	0.9198	0.8849
C4 Median	0.9059	0.9168	0.9024	0.9210	0.8832
C5 Average	0.7134	0.8062	0.7793	0.7598	0.6565
C5 Product	0.3644	0.4516	0.5463	0.6272	0.6244
C5 Minimum	0.7100	0.7955	0.7544	0.7297	0.6334
C5 Maximum	0.7137	0.8080	0.7845	0.7683	0.6728
C5 Median	0.7135	0.8066	0.7804	0.7624	0.6603

Table C.3: People detection performance fusing the five detectors (without Fusion detector [Fernández-Carbajales et al., 2008]), in terms of area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
Average	0.8729	0.9127	0.8875	0.8807	0.8019
Product	0.4529	0.6015	0.7139	0.8065	0.7998
Minimum	0.8720	0.9099	0.8812	0.8731	0.8017
Maximum	0.8729	0.9020	0.8864	0.8716	0.8010
Median	0.8729	0.9019	0.8856	0.8709	0.7986

Table C.4: Total average fusion performance fusing the five detectors (without Fusion detector [Fernández-Carbajales et al., 2008]), in terms of area under the Precision-Recall curve (AUC-PR) of dataset A.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
C5 Average	0.5491	0.6497	0.6921	0.6637	0.6256	0.5861
C5 Product	0.4542	0.5960	0.6595	0.6545	0.6203	0.5847
C5 Minimum	0.5468	0.6442	0.6749	0.6496	0.6101	0.5689
C5 Maximum	0.5344	0.6376	0.6802	0.6651	0.6406	0.5977
C5 Median	0.5392	0.6354	0.6735	0.6660	0.6230	0.5944

Table C.5: People detection performance fusing the six appearance based detectors, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
C5 Average	0.6224	0.7045	0.7054	0.6651	0.6043
C5 Product	0.4697	0.6173	0.6661	0.6403	0.5985
C5 Minimum	0.6110	0.6891	0.6808	0.6521	0.5859
C5 Maximum	0.6147	0.6939	0.6982	0.6664	0.6223
C5 Median	0.6078	0.7007	0.6919	0.6622	0.6075

Table C.6: People detection performance fusing the five appearance based detectors (without Fusion detector [Fernández-Carbajales et al., 2008]), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B.

	$m=1$	$m=2$	$m=3$
C5 Average	0.6986	0.7224	0.6451
C5 Product	0.5834	0.6771	0.6369
C5 Minimum	0.6963	0.7141	0.6403
C5 Maximum	0.6953	0.6984	0.6420
C5 Median	0.6939	0.7083	0.6453

Table C.7: People detection performance fusing the three appearance based detectors (HOG, ISM and DTDP), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
C5 Average	0.5441	0.6447	0.6819	0.6637	0.6306	0.5911
C5 Product	0.4542	0.6010	0.6545	0.6495	0.6203	0.5797
C5 Minimum	0.5418	0.6392	0.6699	0.6496	0.6151	0.5739
C5 Maximum	0.5344	0.6376	0.6802	0.6651	0.6356	0.5977
C5 Median	0.5392	0.6404	0.6785	0.6610	0.6280	0.5894

Table C.8: People detection performance fusing the six appearance based detectors, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
C5 Average	0.6174	0.6996	0.7004	0.6701	0.6093
C5 Product	0.4747	0.6123	0.6611	0.6453	0.5935
C5 Minimum	0.6110	0.6891	0.6858	0.6521	0.5909
C5 Maximum	0.6097	0.6939	0.6982	0.6714	0.6173
C5 Median	0.6128	0.6957	0.6969	0.6672	0.6075

Table C.9: People detection performance fusing the five appearance based detectors (without Fusion detector [Fernández-Carbajales et al., 2008]), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

	$m=1$	$m=2$	$m=3$
C5 Average	0.6986	0.7131	0.6501
C5 Product	0.5784	0.6821	0.6319
C5 Minimum	0.6963	0.7091	0.6403
C5 Maximum	0.6903	0.7034	0.6420
C5 Median	0.6939	0.7083	0.6453

Table C.10: People detection performance fusing the three appearance based detectors (HOG, ISM and DTDP), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
C5 Average	0.5625	0.6601	0.6987	0.6770	0.6407	0.5947
C5 Product	0.4689	0.6108	0.6678	0.6630	0.6300	0.5871
C5 Minimum	0.5601	0.6560	0.6900	0.6684	0.6322	0.5881
C5 Maximum	0.5532	0.6517	0.6941	0.6740	0.6394	0.5927
C5 Median	0.5577	0.6555	0.6947	0.6734	0.6373	0.5918

Table C.11: People detection performance fusing the six appearance and motion based detectors combinations, in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
C5 Average	0.6343	0.7182	0.7165	0.6773	0.6108
C5 Product	0.4877	0.6234	0.6802	0.6570	0.6005
C5 Minimum	0.6287	0.7117	0.7086	0.6680	0.6043
C5 Maximum	0.6258	0.7105	0.7109	0.6742	0.6111
C5 Median	0.6297	0.7141	0.7125	0.6738	0.6076

Table C.12: People detection performance fusing the five appearance and motion based detectors combinations (without Fusion+IMM detector), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

	$m=1$	$m=2$	$m=3$
C5 Average	0.7150	0.7268	0.6439
C5 Product	0.5870	0.6869	0.6259
C5 Minimum	0.7126	0.7181	0.6370
C5 Maximum	0.7061	0.7113	0.6356
C5 Median	0.7100	0.7158	0.6389

Table C.13: People detection performance fusing the three appearance and motion based detectors combinations (HOG+IMM, ISM+IMM and DTDP+IMM), in terms of area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B with motion.

Appendix D

Publications

The following publications have been produced in association with this thesis (listed by chapters):

- People detection benchmarking framework (chapter 3)
  - A. García-Martín, J. M. Martínez, J. Bescós. A corpus for benchmarking of people detection algorithms. *Pattern Recognition Letters*, Volume 33, Issue 2, January 2012, Pages 152-156, ISSN 0167-8655, `http://dx.doi.org/10.1016/j.patrec.2011.09.038`.
  - J. C. SanMiguel, A. García-Martín, J. M. Martínez. Performance evaluation in video-surveillance systems: the EventVideo project evaluation protocols. *Intelligent Multimedia Surveillance: Current Trends and Research*, Pradeep Atrey, Mohan Kankanhalli, Andrea Cavallaro (eds.), 2013, Springer (in press).
- Real-time people detection based on appearance information (chapter 4)
  - A. García-Martín, J. M. Martínez. Robust Real Time Moving People Detection in Surveillance Scenarios. *In Proc. of the IEEE International Conference on Advanced Video and Signal based Surveillance*, Pages 241-247, Boston (MA, USA), 29-1 August-September 2010, `http://dx.doi.org/10.1109/AVSS.2010.33`.
- People detection based on appearance and motion information (chapter 5)
  - A. García-Martín, A. Hauptmann, J. M. Martínez. People detection based on appearance and motion models. *In Proc. of the IEEE International Conference on Advanced Video and Signal based Surveillance*, Pages 256-260, Klagenfurt (Austria), 30-2 August-September 2011, `http://dx.doi.org/10.1109/AVSS.2011.6027333`.
- Collaborative people detection and tracking (chapter 6)

  - A. García-Martín, J. M. Martínez. On collaborative people detection and tracking in complex scenarios. *Image and Vision Computing*, Volume 30, Issues 4–5, May 2012, Pages 345-354, ISSN 0262-8856, `http://dx.doi.org/10.1016/j.imavis.2012.03.005`.
  - A. García-Martín, J. M. Martínez. Enhanced people detection combining appearance and motion information. *Electronic Letters*, Volume 49, Issue 4, January 2013, Pages 256-258, ISSN 0013-5194.

- People-background segmentation (chapter 7)
  - A. García-Martín, A. Cavallaro, J. M. Martínez. People-background segmentation with unequal error cost. *In Proc. of the IEEE International Conference on Image Processing*, Orlando (FL, USA), 30-3 September-October 2012.

Apéndice E

Logros, conclusiones y trabajo futuro

E.1 Resumen de logros y principales conclusiones

Esta tesis ha estudiado la detección de personas en escenarios de videovigilancia. El objetivo es analizar las aproximaciones más representativas del estado del arte, identificar sus debilidades y proponer contribuciones para mejorar las aproximaciones actuales de detección de personas. En particular, se han explorado dos áreas relacionadas con la evaluación y comparación de algoritmos de detección de personas (capítulo 3) y aproximaciones a la detección de personas (capítulos 4, 5, 6, 7 y 8).

En la primera parte de esta tesis, hemos descrito las motivaciones y consideraciones necesarias en el diseño y generación de un conjunto de vídeos o corpus (vídeos y anotaciones asociadas) y la definición de una metodología de evaluación de algoritmos de detección de personas en secuencias de vídeo (capítulo 3). Se ha producido un conjunto de vídeos más completo que los disponibles actualmente en el estado del arte para la detección de personas en escenarios de videovigilancia (PDds). El amplio número de factores críticos considerados durante el diseño del corpus y la disponibilidad de las correspondientes anotaciones, hacen a nuestro corpus especialmente apto para probar algoritmos, así como para la evaluación y comparación de resultados. Se ha definido una metodología para la evaluación de la detección de personas, con el particular interés de evaluar el funcionamiento o rendimiento global del sistema de detección en vez de evaluar únicamente la tarea de clasificación (persona/no persona). En conjunto, se dispone de un marco común de trabajo para la evaluación de algoritmos de detección de personas bajo diferentes condiciones de complejidad.

En la segunda parte de esta tesis, hemos propuesto tres algoritmos diferentes de detección de personas. En primer lugar, se ha propuesto un detector de personas que combina ambas técnicas de extracción de objetos iniciales candidatos a ser persona, i.e., la segmentación y la búsqueda exhaustiva, con el fin de obtener mayor robustez y ser capaz de operar en tiempo real (capítulo 4). Un sistema completo de videovigilancia ha sido implementado para evaluar el detector prop-

uesto. Además, con el fin de realizar una correcta evaluación del sistema, se ha evaluado sobre nuestro conjunto de vídeos de evaluación PDds. Los resultados obtenidos sobre el conjunto de vídeos A muestran como el sistema propuesto funciona considerablemente bien en tiempo real, funciona incluso mejor que otras propuestas del estado del arte que no operan en tiempo real y es significativamente más eficiente y estable que otras propuestas del estado del arte. Sin embargo, debido a la dificultad de segmentar el fondo de la escena en escenarios complejos, nuestra aproximación obtiene resultados similares a los del estado del arte a niveles altos de complejidad. Los resultados obtenidos sobre el conjunto de vídeos B demuestran que nuestra aproximación no funciona correctamente en escenarios más complejos o realistas. Nuestra propuesta presenta una fuerte dependencia con la etapa de segmentación, por lo que heredamos todos los problemas de la segmentación (segmentación deficiente o sobresegmentación). Nuestra combinación de segmentación y búsqueda exhaustiva reduce estos problemas, pero dichos problemas se ven incrementados en escenarios complejos donde es muy difícil obtener una segmentación fiable.

En segundo lugar, hemos propuesto un detector de personas que combina un modelo de persona del estado del arte basado en apariencia y nuestro modelo de persona basado en movimiento (capítulo 5). Usando el mismo esquema que el ISM y el detector y descriptor de puntos de interés MoSIFT, presentamos un nuevo modelo de persona basado en los movimientos característicos de las personas. Los experimentos se han realizado sobre un conjunto de vídeos de alta complejidad o realistas, extraídos del dataset TRECVID y que forman parte de la categoría de máxima complejidad de nuestro conjunto de vídeos de evaluación PDds. Los resultados muestran como la información de movimiento es muy útil para la detección de las personas e independiente de la información de apariencia, nuestro detector basado en movimiento obtiene resultados comparables a la aproximación del estado del arte ISM en escenarios complejos o realistas. La evaluación del sistema completo demuestra que la combinación de ambas fuentes de información independientes mejora la detección final, obteniendo una mejora significativa del Recall y una ligera reducción de la Precisión.

En tercer lugar, esta tesis ha investigado la posibilidad de aprovechar la combinación de apariencia y movimiento a lo largo del tiempo mediante un sistema colaborativo de detección de personas y su seguimiento (capítulo 6). Se ha integrado la información de detección de personas y su seguimiento en un único sistema que mejora ambas tareas simultáneamente. Hemos analizado las diferentes configuraciones del sistema con el fin de evaluar la mejora introducida por el intercambio mutuo de información entre tareas. Los experimentos se han realizado sobre un conjunto de vídeos de alta complejidad o realistas de nuestro conjunto de vídeos de evaluación PDds y extraídos del dataset TRECVID (escenas altamente pobladas, fondos de alta complejidad y personas a múltiples escalas), destacando los problemas que este tipo de escenarios tan complejos implican en el estado del arte de detección de personas y seguimiento. Los resultados sobre el conjunto de vídeos propuesto demuestran la utilidad de nuestro sistema colaborativo, especialmente

en escenarios complejos, obteniendo mejores resultados que el estado del arte en ambas tareas independientemente. Los módulos de detección y seguimiento pueden ser fácilmente remplazados por otros gracias al diseño modular del sistema que permite tanto el funcionamiento colaborativo como independiente, el formato genérico de la información intercambiada (localización, dimensión y la confianza de la detección/seguimiento) y el altamente compatible mecanismo de intercambio de información (proceso de actualización simple y consistente). El uso de diferentes módulos variará el rendimiento global del sistema, pero la combinación de ambas fuentes de información, en principio, resultará útil para mejorar el sistema (excepto en el caso ideal en el que tengamos un funcionamiento perfecto de las tareas de detección y seguimiento). En relación a la detección de personas, en primer lugar, hemos usado un detector basado en la combinación de información de apariencia y movimiento. Hemos evaluado las diferentes combinaciones de apariencia y movimiento con diferentes detectores del estado del arte y se ha demostrado la utilidad de la información de movimiento y su independencia con la información de apariencia. En segundo lugar, se ha propuesto un esquema de predicción o actualización de la detección de personas usando la información de seguimiento de nuestro sistema colaborativo y se han re-evaluado de nuevo todas las variaciones de detectores de personas. Los resultados experimentales muestran como la información de seguimiento permite estabilizar la detección de personas a lo largo del tiempo, por lo que se traduce en una mejora significativa sobretodo en términos de Recall y F1Score. En relación al seguimiento, en primer lugar, se ha evaluado un algoritmo de seguimiento de filtrado de partículas adaptativo basado en la distribución de color del modelo a seguir y con diferentes inicializaciones usando los diferentes detectores de personas. Todas las diferentes inicializaciones siguen un patrón de comportamiento similar, pero se observa claramente que el proceso de inicialización tiene una gran influencia en el funcionamiento global del seguimiento. En segundo lugar, se ha introducido la información de detección de personas de nuestro sistema colaborativo y se han re-evaluado las diferentes variaciones de seguimiento. Los resultados experimentales muestran como el uso de la información de detección de personas ayuda a corregir la posición, dimensión y por lo tanto la distribución de color utilizada para seguir a cada persona a lo largo del tiempo, por lo que se traduce en una mejora principalmente en términos de Precisión y F1Score.

Durante la segunda parte de la tesis, además de los algoritmos de detección hemos propuesto dos tareas adicionales de post-procesado. En primer lugar, hemos propuesto un segmentador persona-fondo que trata de asegurar que ninguna persona o parte del cuerpo de una persona son asignadas o clasificadas como fondo, a costa de potencialmente incrementar el número de píxeles del fondo clasificados como persona y, entonces, se ha propuesto una nueva tarea de post-procesado basada en esta segmentación persona-fondo (capítulo 7). Los experimentos realizados muestran el funcionamiento de nuestra propuesta sobre el dataset de evaluación propuesto PDds. Se aprecia una mejora global en casi todas las categorías y sobre el funcionamiento original de los

detectores, siendo dicha mejora más evidente en aquellos escenarios con complejidad media o alta del fondo de la escena, ya que estos escenarios son más probables de generar falsas detecciones. En segundo lugar, también se ha propuesto la combinación o fusión de hasta seis detectores de personas independientes basados en apariencia y su combinación con nuestro detector basado en movimiento, con el objetivo de mejorar la detección en escenarios típicos de videovigilancia (capítulo 8). Para poder fusionar los diferentes detectores, se ha presentado un criterio de combinación de múltiples detecciones y la aplicación de técnicas de fusión tradicionales: promedio, producto, mínimo, máximo y mediana. Los experimentos realizados muestran el funcionamiento de nuestra propuesta con cada una de las técnicas de fusión mencionadas. El método de producto muestra claramente los peores resultados, mientras que el método de promedio obtiene resultados ligeramente mejores que los otros tres métodos. Los experimentos realizados también muestran el funcionamiento de nuestra propuesta sobre el dataset de evaluación propuesto PDds. Se aprecia una mejora global en todas las categorías y sobre el funcionamiento original de los detectores, siendo dicha mejora más evidente en aquellos escenarios con mayor complejidad, ya que estos escenarios son más probables de generar falsas detecciones y perdida de detecciones correctas. Finalmente, en ambos casos, los resultados muestran como el uso de la información de movimiento junto con ambas tareas de post-procesado obtienen los mejores resultados finales.

E.2 Trabajo futuro

Basándose en los resultados y conclusiones de esta tesis, se proponen las siguientes extensiones:

- Ampliación del conjunto de vídeos de evaluación PDds. En el capítulo 3, el conjunto de vídeos de evaluación propuesto PDds incluye gran variedad de escenarios con diferentes complejidades de fondo, variedad de apariencia de las personas y múltiples interacciones de personas con objetos y/o otras personas. Sin embargo, proponemos ampliar el número de secuencias de vídeo de evaluación, utilizar todas las secuencias grabadas en un estudio de grabación con un sistema de chroma y componer las múltiples combinaciones de secuencias con cada frente y fondo disponibles [Tiburzi et al., 2008], de esta forma sería posible evaluar de forma independiente los factores de frente y fondo que afectan a la detección.

- Mejorar o refinar la segmentación o sustracción del fondo de la escena. El detector de personas propuesto en el capítulo 4 combina la segmentación y la búsqueda exhaustiva. Tal y como demuestran los resultados experimentales obtenidos, nuestra combinación de segmentación y búsqueda exhaustiva reduce los problemas inherentes a la segmentación (segmentación deficiente o sobresegmentación), pero estos problemas se ven amplificados en escenarios complejos donde es muy difícil obtener una segmentación fiable. Por este motivo, para tratar de mejorar o refinar la segmentación del fondo en escenarios com-

plejos, proponemos el estudio y aplicación de técnicas de modelado del fondo en escenas multimodales, reducción del ruido de segmentación, eliminación de sombras, etc.

- Fusión de información de apariencia y movimiento. En el capítulo 5, hemos propuesto un detector de personas que combina dos detectores independientes: un modelo de persona del estado del arte basado en apariencia (ISM) y nuestro modelo de persona basado en movimiento (IMM). Proponemos estudiar diferentes técnicas de fusión o combinación entre la salida de ambos detectores con el objetivo de tratar de mejorar el Recall sin comprometer la Precisión, o incluso la creación de un único e integrado Implicit Shape-Motion Model (ISMM), usando el descriptor completo MoSIFT.

- Ampliar la evaluación de la tarea de seguimiento. El sistema colaborativo de detección de personas y su seguimiento presentado en el capítulo 6 ha sido evaluado con un único módulo de seguimiento en particular. Sin embargo, el módulo de seguimiento puede ser fácilmente remplazado por otros gracias al diseño modular del sistema que permite tanto el funcionamiento colaborativo como independiente, el formato genérico de la información intercambiada y el altamente compatible mecanismo de intercambio de información. Por lo que proponemos la evaluación del sistema colaborativo con otros algoritmos de seguimiento o incluso, como en el caso de la detección de personas, combinar eficientemente múltiples algoritmos de seguimiento independientes.

- Esquemas colaborativos hacia adelante y atrás. En el capítulo 6, se ha propuesto un sistema colaborativo de detección de personas y su seguimiento hacia adelante. Proponemos no sólo utilizar este esquema colaborativo de hacia adelante, sino también utilizar aproximaciones en ambos sentidos hacia adelante y atrás, i.e., sistemas retroalimentados.

- Segmentación persona-fondo. Proponemos mejorar la segmentación persona-frente presentada en el capítulo 7 incorporando información temporal en el modelo y explorar la posibilidad de detectar automáticamente el rango de escalas de personas presentes en cada parte de la escena así como el umbral de binarización. Además, proponemos extender el método de segmentación propuesto a otros detectores de personas y otras clases de objetos.

- Confianza de la segmentación. Con el objetivo de mejorar la confianza de la segmentación propuesta en el capítulo 7, proponemos la combinación de la segmentación persona-fondo con otras estrategias de segmentación clásicas: basadas en color, movimiento, etc. Tras comprobar que el post-procesado propuesto mejora los resultados, proponemos estudiar el uso del segmentador persona-fondo como una etapa de preprocesado con el objetivo de mantener o reducir el coste computacional. Finalmente, también proponemos explorar otras posibles combinaciones de la detección y la confianza de la segmentación.

- Fusión a nivel de decisión de detectores de personas. En el capítulo 8, se ha propuesto la combinación o fusión de seis detectores de personas independientes basados en apariencia y uno basado en movimiento. Proponemos explorar otros métodos de fusión más complejos, no solo usando reglas fijas, sino también reglas aprendidas o el uso de pesos adaptativos en base a estimaciones online de calidad, además proponemos no sólo usar esquemas de fusión en paralelo, sino también en cascada, jerárquicas o híbridas. Finalmente, está claro que detectores "construidos de forma independiente" presentan correlaciones entre ellos y esto se debe al hecho de que existen zonas del espacio de decisión difíciles para todos los detectores. Por lo tanto, también proponemos explorar otras técnicas de fusión que sean robustas a la correlación de decisores.

Glossary

ANP *Average Number of People*

AP *Average Precision*

AUC *Area under the curve*

AUC-PR *Area under the Precision-Recall curve*

AUC-ROC *Area under the Receiver Operating Characteristics curve*

AVSS *Advanced Video and Signal-Based Surveillance*

BLOB *Binary Large Object*

CPU *Central Processing Unit*

DBP *Dependent Body Parts*

DEBP *Dependent Extended Body Parts*

DEBP-P *Dependent Extended Body Parts Post-processed*

DET *Detection Error Tradeoff*

DTDP *Discriminatively Trained Deformable Parts*

FPR *False Positive Rate*

GTF *Ground-Truth Frames*

HOG *Histogram of Oriented Gradients*

IBP *Independent Body Parts*

IEBP *Independent Extended Body Parts*

IMM *Implicit Motion Model*

ISM *Implicit Shape Model*

ISMM *Implicit Shape-Motion Model*

LBP *Local Binary Patterns*

MHSC *Multiple Hypotheses Simplification Criteria*

MMI *Maximum Motion Information*

MoSIFT *Motion Scale-Invariant Feature Transform*

PASCAL *Pattern Analysis, Statistical Modelling and Computational Learning*

PDds *Person Detection dataset*

PETS *Performance Evaluation of Tracking and Surveillance*

PPP *Percentage of Pixels of People*

PR *Precision-Recall*

RAM *Random Access Memory*

RNN *Reciprocal Nearest Neighbors*

ROC *Receiver Operating Characteristics*

ROI *Region Of Interest*

SIFT *Scale-Invariant Feature Transform*

SVM *Support Vector Machine*

TPR *True Positive Rate*

TRECVID *TREC Video Retrieval Evaluation*

TUD *Technische Universität Darmstadt*

URL *Universal Resource Locator*

ViPER *Video Performance Evaluation Resource*

XML *eXtensible Markup Language*

Bibliography

I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocaña, and M. A. G. Garrido. Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):292–307, 2007. [Cited on pages 12, 13, 14, 15, 16, 18, 19, and 40.]

M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. of CVPR*, pages 1–8, 2008. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 24, 25, 34, 44, 51, 63, 64, 78, and 121.]

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. of CVPR*, pages 1014–1021, 2009. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 40, 44, 52, 70, 78, 90, 99, 101, 119, and 120.]

S. Avidan. Ensemble tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2): 261–271, 2007. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

AVSS. International conference on advanced video and signal based surveillance, http://www.avss2007.org. [Cited on pages 24 and 28.]

M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1820–1833, 2010. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, 78, and 98.]

T. J. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2):90–99, 1986. [Cited on page 41.]

A. Cavallaro and T. Ebrahimi. Video object extraction based on adaptive background and statistical change detection. In *Proc. of SPIE*, pages 465–475, 2001. [Cited on page 41.]

M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University, 2009. [Cited on pages 52, 55, and 56.]

D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003. [Cited on page 62.]

D. Cremers and C. Schnörr. Statistical shape knowledge in variational motion segmentation. *Image and Vision Computing*, 21(1):77–86, 2003. [Cited on page 62.]

X. Cui, Y. Liu, S. Shan, X. Chen, and W. Gao. 3d haar-like features for pedestrian detection. In *Proc. of ICME*, pages 1263–1266, 2007. [Cited on pages 12, 13, 14, 16, 18, 19, and 20.]

R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000. [Cited on pages 12, 13, 14, 15, 17, 18, 19, 26, 52, and 55.]

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, pages 886–893, 2005. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 24, 25, 32, 40, 44, 45, 52, 70, 78, 83, 90, 98, 99, 100, 119, and 120.]

N. Dalal and B. Triggs. Human detection using oriented histograms of flow and appearance. In *Proc. of ECCV*, pages 428–441, 2006. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 52, 55, and 98.]

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proc. of ICML*, pages 233–240, 2006. [Cited on page 34.]

S. Denman, V. Chandran, and S. Sridharan. An adaptive optical flow technique for person tracking systems. *Pattern Recognition Letters*, 28(10):1232–1239, 2007. [Cited on page 62.]

M. B. Dillencourt, H. Samet, and M. Tamminen. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM*, 39(2):253–280, 1992. [Cited on page 41.]

P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. of CVPR*, pages 304–311, 2009. [Cited on pages 24, 25, and 26.]

P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. [Cited on pages 9 and 32.]

M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009. [Cited on pages 9, 24, 25, and 32.]

A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *Proc. of ICCV*, pages 1–8, 2007. [Cited on pages 24 and 25.]

A. Ess, B. Leibe, K. Schindler, and L. V. Gool. A mobile vision system for robust multi-person tracking. In *Proc. of CVPR*, pages 1–8, 2008. [Cited on page 122.]

A. Ess, B. Leibe, K. Schindler, and L. V. Gool. Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1831–1846, 2009. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, 78, and 82.]

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 40, 44, 70, 78, 83, 84, 85, 86, 90, 95, 98, 99, 101, 119, 120, 121, 122, 123, 124, and 125.]

V. Fernández-Carbajales, M. A. García, and J. M. Martínez. Robust people detection by fusion of evidence from multiple methods. In *Proc. of WIAMIS*, pages 55–58, 2008. [Cited on pages 12, 13, 14, 15, 18, 19, 44, 90, 98, 99, 100, 102, 103, 105, 106, 119, 132, 133, and 134.]

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computer and System Sciences*, 55(1):119–139, 1997. [Cited on pages 42 and 43.]

G. Gan and J. Cheng. Pedestrian detection based on hog-lbp feature. In *Proc. of CIS*, pages 1184–1187, 2011. [Cited on page 98.]

D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007. [Cited on pages 12, 13, 14, 15, 18, 19, and 40.]

D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7): 1239–1258, 2010. [Cited on pages 9 and 82.]

J. Giebel, D. M. Gavrila, and C. Schnorr. A bayesian framework for multi-cue 3d object tracking. In *Proc. of ECCV*, pages 241–252, 2004. [Cited on pages 12, 13, 14, 15, 18, 19, 20, 63, 64, and 78.]

N. Haering, P. L. Venetianer, and A. Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5-6):279–290, 2008. [Cited on page 3.]

D. Hall and J. Llinas. *Handbook of multisensor data fusion*. Electrical Engineering & Applied Signal Processing Series. CRC Press Inc, June 2001. [Cited on page 98.]

S. Harasse, L. Bonnaud, and M. Desvignes. Human model for people detection in dynamic scenes. In *Proc. of CVPR*, pages 335–354, 2006. [Cited on pages 12, 13, 14, 15, 18, and 19.]

I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost: a human body part labeling system using silhouettes. In *Proc. of ICPR*, pages 77–82, 1998. [Cited on pages 52 and 119.]

I. Haritaoglu, D. Harwood, and L. S. Davis. W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000. [Cited on pages 12, 13, 14, 15, 18, and 19.]

W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004. [Cited on pages 10 and 26.]

C. Huang, H. Al, B. Wu, and S. Lao. Boosting nested cascade detector for multi-view face detection. In *Proc. of ICPR*, pages 415–418, 2004. [Cited on page 42.]

M. Hussein, W. Abd-Almageed, Y. Ran, and L. Davis. Real-time human detection, tracking, and verification in uncontrolled camera motion environments. In *Proc. of ICVS*, pages 41–47, 2006. [Cited on pages 12, 13, 14, 15, 18, and 19.]

A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, 2003. [Cited on page 62.]

P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110(1):43–59, 2008. [Cited on pages 12, 13, 14, 15, 18, 19, and 26.]

N. Koenig. Toward real-time human detection and tracking in diverse environments. In *Proc. of ICDL*, pages 94–98, 2007. [Cited on pages 12, 13, 14, 15, 18, and 19.]

L. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002. [Cited on pages 98, 99, and 103.]

J. J. Lee. Libpmk: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory, April 2008. [Cited on page 58.]

B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *Proc. of DAGM*, pages 145–153, 2004. [Cited on pages 12, 13, 14, 16, 18, 19, 20, and 40.]

B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. of CVPR*, pages 878–885, 2005. [Cited on pages 34, 35, 44, 51, 56, 66, 71, 90, 99, 101, 119, and 120.]

B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *Proc. of ICCV*, pages 1–8, 2007. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 34, 52, 53, 54, 55, 56, 59, 70, and 78.]

Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1728–1740, 2008. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Proc. of DAGM*, pages 297–304, 2003. [Cited on page 43.]

D. Lowe. Distinctive image features from scale invariant key points. *International Journal of Computer Vision*, 60(2):91–110, 2004. [Cited on page 55.]

S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006. [Cited on pages 24, 25, and 32.]

A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proc. of AVSS*, pages 476–481, 2007. [Cited on page 24.]

K. Nummiaro, E. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21(1):99–110, 2003. [Cited on pages 53, 62, 66, 68, 71, and 78.]

T. Ojala and M. Pietikainen. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7): 971–988, 2002. [Cited on page 98.]

K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. of ECCV*, pages 28–39, 2004. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000. [Cited on pages 24 and 25.]

PASCAL. The pascal visual object classes challenge 2005 development kit, http://pascallin.ecs.soton.ac.uk. [Cited on page 25.]

PETS. International workshop on performance evaluation of tracking and surveillance, http://www.pets2006.net/. [Cited on pages 24 and 28.]

K. N. Plataniotis and C. S. Regazzoni. Visual-centric surveillance networks and services. *IEEE Signal Processing Magazine*, 22(2):12–15, 2005. [Cited on page 3.]

P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, 2004. [Cited on page 62.]

C. S. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, and A. Hampapur. Video analytics for surveillance: Theory and practice. *IEEE Signal Processing Magazine*, 27(5):16–17, 2010. [Cited on page 3.]

X. Ren. Finding people in archive films through tracking. In *Proc. of CVPR*, pages 1–8, 2008. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008. [Cited on page 45.]

E. Seemann and B. Schiele. Cross-articulation learning for robust detection of pedestrians. In *Proc. of DAGM*, pages 242–252, 2006. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 40, and 44.]

H. Sidenbladh. Detecting human motion with support vector machines. In *Proc. of ICPR*, pages 188–191, 2004. [Cited on pages 12, 13, 14, 16, 17, 18, 19, and 52.]

N. Sprague and J. Luo. Clothed people detection in still images. In *Proc. of ICPR*, pages 585–589, 2002. [Cited on pages 12, 13, 14, 15, 18, and 19.]

S. Stalder, H. Grabner, and L. V. Gool. Cascaded confidence filtering for improved tracking-by-detection. In *Proc. of ECCV*, pages 369–382, 2010. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

F. Tiburzi, M. Escudero, J. Bescós, and J. M. Martínez. A ground truth for motion-based video-object segmentation. In *Proc. of ICIP*, pages 17–20, 2008. [Cited on pages 26, 28, 115, and 142.]

TRECVID. Trecvid 2008 evaluation for surveillance event detection, http://www-nlpir.nist.gov/projects/trecvid/. [Cited on pages 24, 28, and 32.]

M. Valera and S. A. Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings on Visual Image Signal Processing*, 152(2):192–204, 2005. [Cited on pages 3, 10, 26, and 41.]

R. Vezzani and R. Cucchiara. Annotation collection and online performance evaluation for video surveillance: The visor project. In *Proc. of AVSS*, pages 227–234, 2008. [Cited on pages 24 and 28.]

P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2): 137–154, 2004. [Cited on pages 12, 13, 14, 16, 18, and 19.]

P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, pages 511–518, 2001. [Cited on pages 41 and 43.]

P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. of ICCV*, pages 734–741, 2003. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 40, 52, 55, and 98.]

ViPER. Viper-gt, the ground truth authoring tool, http://vipertoolkit.sourceforge.net/docs/gt/. [Cited on pages 25 and 28.]

J. Wang and Y. Yagi. Integrating color and shape-texture features for adaptive real-time object tracking. *IEEE Transactions on Image Processing*, 17(2):235–240, 2008. [Cited on page 62.]

C. Wojek, G. Dorkó, A. Schulz, and B. Schiele. Sliding-windows for rapid object class localization: A parallel technique. In *Proc. of DAGM*, pages 71–81, 2008. [Cited on pages 12, 13, 14, 16, 18, 19, 20, and 40.]

C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Proc. of CVPR*, pages 794–801, 2009. [Cited on pages 24, 25, 34, and 51.]

B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. of ICCV*, pages 90–97, 2005. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 26, 39, 40, 41, 42, 43, 44, 45, 52, 98, and 119.]

B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 40, 44, 63, 64, and 78.]

F. Xu and K. Fujimura. Human detection using depth and gray images. In *Proc. of AVSS*, pages 115–121, 2003. [Cited on pages 12, 13, 14, 15, 18, 19, 26, 52, and 119.]

A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):1–45, 2006. [Cited on page 62.]

J. Yu, D. Farin, and B. Schiele. Multi-target tracking in crowded scenes. In *Proc. of DAGM*, pages 406–415, 2011. [Cited on pages 12, 13, 14, 16, 18, 19, 20, 63, 64, and 78.]

W. Zhang, G. Zelinsky, and D. Samaras. Real-time accurate object detection using multiple resolutions. In *Proc. of ICCV*, pages 1–8, 2007. [Cited on pages 12, 13, 14, 16, 18, 19, 20, and 40.]

T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, 2004. [Cited on pages 12, 13, 14, 15, 18, 19, and 40.]

H. Zhou, Y. Yuan, and C. Shi. Object tracking using sift features and mean shift. *Computer Vision and Image Understanding*, 113(3):345–352, 2009. [Cited on page 62.]

J. Zhou and J. Hoang. Real time robust human detection and tracking system. In *Proc. of CVPR*, pages 149–156, 2005. [Cited on pages 12, 13, 14, 15, 18, 19, 40, and 52.]

Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. of CVPR*, pages 1491–1498, 2006. [Cited on pages 12, 13, 14, 16, 18, 19, 20, and 40.]